Photometric Reconstruction from Images: New Scenarios and Approaches for Uncontrolled Input Data

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

DISSERTATION

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
von

Diplom-Mathematiker Jens Ackermann
geboren in Gießen.

Referenten der Arbeit:	Prof. Dr.-Ing. Michael Goesele Technische Universität Darmstadt
	Prof. Dr. Reinhard Klein Rheinische Friedrich-Wilhelms-Universität Bonn
Tag der Einreichung:	23.04.2014
Tag der mündlichen Prüfung:	16.06.2014

Darmstadt 2014
D17

Erklärung zur Dissertation

Hiermit versichere ich die vorliegende Dissertation selbständig nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 23.04.2014 Jens Ackermann

Abstract

The changes in surface shading caused by varying illumination constitute an important cue to discern fine details and recognize the shape of textureless objects. Humans perform this task subconsciously, but it is challenging for a computer because several variables are unknown and intermix in the light distribution that actually reaches the eye or camera. In this work, we study algorithms and techniques to automatically recover the surface orientation and reflectance properties from multiple images of a scene.

Photometric reconstruction techniques have been investigated for decades but are still restricted to industrial applications and research laboratories. Making these techniques work on more general, uncontrolled input without specialized capture setups has to be the next step but is not yet solved. We explore the current limits of photometric shape recovery in terms of input data and propose ways to overcome some of its restrictions.

Many approaches, especially for non-Lambertian surfaces, rely on the illumination and the radiometric response function of the camera to be known. The accuracy such algorithms are able to achieve depends a lot on the quality of an a priori calibration of these parameters. We propose two techniques to estimate the position of a point light source, experimentally compare their performance with the commonly employed method, and draw conclusions which one to use in practice. We also discuss how well an absolute radiometric calibration can be performed on uncontrolled consumer images and show the application of a simple radiometric model to re-create night-time impressions from color images.

A focus of this thesis is on Internet images which are an increasingly important source of data for computer vision and graphics applications. Concerning reconstructions in this setting we present novel approaches that are able to recover surface orientation from Internet webcam images. We explore two different strategies to overcome the challenges posed by this kind of input data. One technique exploits orientation consistency and matches appearance profiles on the target with a partial reconstruction of the scene. This avoids an explicit light calibration and works for any reflectance that is observed on the partial reference geometry. The other technique employs an outdoor lighting model and reflectance properties represented as parametric basis materials. It yields a richer scene representation consisting of shape and reflectance. This is very useful for the simulation of new impressions or editing operations, *e.g.* relighting. The proposed approach is the first that achieves such a reconstruction on webcam data. Both presentations are accompanied by evaluations on synthetic and real-world data showing qualitative and quantitative results.

We also present a reconstruction approach for more controlled data in terms of

the target scene. It relies on a reference object to relax a constraint common to many photometric stereo approaches: the fixed camera assumption. The proposed technique allows the camera and light source to vary freely in each image. It again avoids a light calibration step and can be applied to non-Lambertian surfaces.

In summary, this thesis contributes to the calibration and to the reconstruction aspects of photometric techniques. We overcome challenges in both controlled and uncontrolled settings, with a focus on the latter. All proposed approaches are shown to operate also on non-Lambertian objects.

Zusammenfassung

Als Menschen nutzen wir unbewusst die Helligkeitsverläufe, die durch sich verändernde Beleuchtung hervorgerufen werden, um feine Oberflächendetails zu erkennen oder die Form texturloser Objekte einzuschätzen. Für einen Computer sind solche Aufgaben jedoch sehr herausfordernd, da das Licht, welches unser Auge oder eine Kamera erreicht, durch verschiedene Faktoren bestimmt wird, die sich gegenseitig beeinflussen. In dieser Arbeit untersuchen wir Algorithmen und Methoden, die es ermöglichen Form und Reflektanz von Objekten allein aus Bildern zu rekonstruieren.

Solche photometrischen Methoden sind bereits seit Jahrzehnten Gegenstand der Forschung. Ihre Anwendung beschränkt sich bisher jedoch auf industrielle Umgebungen und Forschungslabors. Der nächste logische Schritt ist daher eine Erweiterung dieser Techniken auf allgemeine, unkontrollierte Eingabedaten, die ohne spezielle Versuchsaufbauten auskommen. Dieses Problem ist in seiner Allgemeinheit noch nicht gelöst. Wir untersuchen daher die Grenzen derzeitiger Verfahren in Bezug auf ihre Eingabedaten und zeigen Wege auf, um einige der existierenden Beschränkungen zu überwinden.

Viele Ansätze, insbesondere für nicht diffuse Oberflächen, beruhen auf der Annahme, dass die Beleuchtung und die radiometrischen Eigenschaften der Kamera bekannt sind. Die Genauigkeit solcher Algorithmen hängt stark von der Qualität einer vorherigen Kalibrierung dieser Parameter ab. Wir schlagen zwei Methoden zur Positionsbestimmung einer Punktlichtquelle vor, vergleichen ihre Leistungsfähigkeit gegenüber der allgemein verbreiteten Vorgehensweise und ziehen daraus Schlüsse, welche in der Praxis zu bevorzugen ist. Außerdem erörtern wir, zu welchem Grad eine absolute radiometrische Kalibrierung auf unkontrollierten Alltagsbildern möglich ist und zeigen, wie es die Anwendung eines simplen radiometrischen Modells erlaubt, aus Farbfotos einer nächtlichen Szene die Eindrücke eines tatsächlichen Beobachters zu simulieren.

Ein Schwerpunkt dieser Arbeit liegt auf Bildern aus dem Internet. Diese werden zu einer immer wichtigeren Quelle für “Computer Vision”- und “Computer Grafik”-Anwendungen. Wir stellen neue Rekonstruktionsverfahren in diesem Umfeld vor. Diese ermöglichen es beispielsweise die Orientierung von Oberflächen in Bildern von Internet-Webcams abzuschätzen. Wir erforschen dabei zwei verschiedene Ansätze, um die Herausforderungen in solchen Daten zu überwinden. Das erste Verfahren nutzt das Konzept der Orientierungskohärenz und findet übereinstimmende Intensitätsprofile zwischen dem Zielobjekt und einer partiellen Rekonstruktion der Szene. Auf diese Weise wird eine explizite Kalibrierung der Beleuchtung vermieden. Darüber hinaus ist das Konzept für beliebige Reflektanzeigenschaften, die in der partiellen Rekonstruktion enthalten sind, anwendbar. Das zweite Verfahren verwendet ein Outdoor-Beleuchtungsmodell und repräsentiert Reflektanzeigenschaften mittels parametrisier-

ter Basismaterialien. Dies resultiert in einer erweiterten Szenenrepräsentation, welche sowohl Form als auch Reflektanz berücksichtigt. Solch eine Darstellung ist insbesondere nützlich, um diese Eigenschaften gezielt zu bearbeiten oder um synthetisch neue Eindrücke der Szene zu erzeugen, z.B. beim "relighting". Der hier vorgeschlagene Ansatz ist der erste der eine derartige Rekonstruktion aus Webcam-Bildern erlaubt. Die Präsentation beider Verfahren wird ergänzt durch qualitative und quantitative Evaluierungen sowohl auf synthetischen als auch auf echten Daten.

Des Weiteren entwickeln wir einen neuen Ansatz für kontrollierte Bedingungen. Dieser beruht auf einem Referenzobjekt und hebt eine verbreitete Einschränkung vieler photometrischer Verfahren auf: die Annahme einer festen Kameraposition. Die vorgeschlagene Methode erlaubt es hingegen sowohl die Kamera als auch die Lichtquelle in jedem Bild frei zu bewegen. Auch diese Technik vermeidet eine Lichtkalibrierung und ist für Lambert'sche wie auch für nicht Lambert'sche Oberflächen anwendbar.

Insgesamt liegen die Beiträge dieser Arbeit im Bereich der Kalibrierung und in der Rekonstruktion mittels photometrischer Methoden. Wir adressieren dabei Herausforderungen sowohl in kontrollierten als auch in unkontrollierten Umgebungen—wobei der Fokus auf letzteren liegt. Außerdem zeigen wir die Anwendbarkeit aller präsentierten Ansätze auch für nicht Lambert'sche Objekte und erweitern so das Spektrum an zulässigen Szenen.

Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr.-Ing. Michael Goesele for his emotional and scientific support, the work environment he created, and his valuable feedback. His trust and openness have contributed a lot to make witnessing the growth of a small team into a full research group an exciting experience.

I would also like to thank Prof. Dr. Reinhard Klein who kindly agreed to review this thesis. His enthusiasm at presenting the best papers at the IGD computer graphics evening is inspiring. Furthermore, I thank Prof. Dr. Wolfgang Heidrich and his group who accommodated me during my research visit at the University of British Columbia in 2011. Prof. Dr. Kay Hamacher and his Computational Biology and Simulation group at the Technische Universität Darmstadt have my gratitude for the collaboration in an interdisciplinary project and many opportunities to discuss topics beyond computer vision. Many thanks to Dr. Philipp Urban for suggestions and support related to camera calibration and color perception. I would also like to thank all external guests at our group retreats for their feedback and suggestions.

I have also benefited from the research ecosystem established by the IGD. This was particularly helpful for all of the structured light scans used as ground truth in this thesis. I am grateful for this support. I especially thank Reiner Weber for his quick response to any technical problem and his help with building some of the capture setups.

Special thanks go to all my colleagues in GRIS who have accompanied me for the last years. Their support in all matters, the cooperation in several projects, the feedback from so many discussions, and bearing with me in stressful times were an important contribution. This includes several visiting researchers from Spain and Austria who brought new perspectives into our group. It also includes the students working in our lab who are an important part of the group and are always ready to help out.

Contents

Chapter 1

Introduction

Since its popularization in the early 20th century, photography has enabled us to capture and *store* our perception of the visual world. On the other hand, modern computer graphics allow us to *create* realistic impressions given a suitable description of a scene. Photographs have the disadvantage of capturing just a single perspective. We observe the world through a window and thus loose some of the impressions an actual observer would have. Computer graphics in connection with virtual reality can create a higher degree of immersion and allows us to be "in the scene". The drawback is that the required scene description is hard to obtain.

How a human observer perceives a scene depends on various factors. Almost all of the light reaching our eyes does not stem directly from a light source but has been reflected at a surface in the scene. Without these reflections, our world would appear mostly dark. Furthermore, the way an object reflects light defines its color and hints at its material, *e.g.*, metallic or plastic. Thus, reflection properties of a surface contain important cues for our understanding of a scene and our perception of the world. The amount of light reaching the eye also depends on the orientation of the surface: specular highlights will move depending on the viewing direction, and a matte plane will appear brighter if oriented towards a source of light. Accordingly, a common way to describe the appearance of a scene is by way of the geometry, its reflectance properties, and the incident lighting.

Creating such an enhanced scene description usually requires a graphics artist or professional. Thus, benefits such as simulating novel impressions or perspectives synthetically are not widely available or applicable to everyday scenes. Digital cameras on the other hand are inexpensive, easy to use, and make photography available as a common tool to preserve the impression of a scene. Image-based reconstruction techniques recover the constituents of scene appearance from one or several images and thus help to bridge the gap between traditional photography and computer graphics.

Such techniques face, however, considerable challenges. Jointly recovering reflectance, shape, and lighting from images has the problem that all of them influence the final brightness of a camera's sensor elements. Observing a surface that reflects only little light in general but is oriented towards the light source can produce the same impression as a surface reflecting more light but oriented at an angle. Most reconstruction techniques therefore assume that at least one of these constituents is known, *e.g.* [Marschner98], and that the others can be approximated with simplified models.

In the last decades those approaches have been shown to yield respectable results for controlled conditions as in a research laboratory. It is, however, unclear how such assumptions generalize for uncontrolled settings and images that have not been captured by an expert. The open questions are more about finding suitable formulations that hold for general input data instead of solving the basic problem or improving its accuracy. In the same spirit, this thesis explores the current limits of what is possible in orientation and reflectance recovery in terms of input data and capturing scenarios.

1.1 Problem Statement

The ultimate goal of computer vision was formulated by Horn in the 1980s:

> "A truly general-purpose vision system would have to deal with all aspects of vision and be applicable to all problems that can be solved using visual information." (B.K.P. Horn [Horn86a])

In our case, we focus on one aspect, image-based reconstruction of scene appearance, and interpret the "applicability to all problems" mostly in terms of the generality of input data. Thus, the goal of reconstruction algorithms should be to recover any kind of object—no matter the shape or reflectance—from arbitrary images, *i.e.* not restricted to a certain illumination, camera model, capture setup, *etc.* After three decades of research this goal is not yet reached. Digital cameras enable anybody to *capture* a scene, but in order to make *reconstructing* the scene common-place more general approaches have to be developed.

An image is the result of the interplay of light and matter in the scene. The goal of *early vision*, according to Bertero *et al.* [Bertero88], is to factorize this combination into its individual components. Thus, the problems discussed in this thesis are so called *inverse problems*. Given a set of rules that transform input data, *e.g.* forces acting on a rigid body or light entering a scene, into output values, *e.g.* velocity change or light arriving at the eye, an inverse problem can in general be formulated as recovering the input when only the output is known. Such tasks can also be interpreted as fitting the parameters of a model to the observed data and arise in many disciplines such as physics, tomography, chemistry, and seismology. They are often ill-posed and may lead to ambiguities if not sufficiently constrained. Furthermore, even slight (random) variations in the observed measurements can lead to large errors in the computed result.

Well-behaved input data is crucial for inverse problems and, accordingly, many image-based reconstruction approaches assume a controlled capture setup. These requirements constrain the application of such techniques mostly to the research community and expensive movie productions, *e.g.* [Alexander10]. This thesis has a strong focus on simple capture setups and uncontrolled input data. We investigate how far the limits can be pushed in this respect and use images downloaded from the Internet as the "ultimate" source of uncontrolled data. While we try to encompass as many components of scene appearance as possible, the emphasis will be on surface shape with repeated consideration of non-Lambertian reflectance.

For image-based shape recovery, different techniques exist such as shape from defocus [Pentland87], shape from texture [Blostein89], or (multi-view) stereo [Seitz06].

These techniques use cues such as the blur induced by finite apertures, the foreshortening of surface patches observed at an angle, or disparity resulting from parallax. In this thesis, we direct our attention to what we call *photometric approaches*. These exploit the intensity variations due to illumination changes. The information encoded in these changes can also be used to recover reflectance properties or the lighting itself.

As mentioned before, photometric approaches usually assume one or two components of scene appearance to be known. In practice, these have to be obtained through a calibration step. A second calibration is required for the measuring device, the camera, whose characteristics can greatly influence the impression of an image and the working of reconstruction algorithms. Thus, for real applications, we have to consider a whole pipeline consisting of one or more preprocessing steps to fulfill all the requirements of the model and the actual reconstruction phase.

In this thesis, we cover several parts of the calibration and at the same time present novel photometric reconstruction techniques that either work on top of these or require (almost) no calibration at all. Two recurring questions are therefore: "To which extent is camera calibration possible on Internet data?" and "What are the factors in traditional approaches that prevent an immediate application to uncontrolled images, and how can these be overcome?".

1.1.1 Internet Data

The vast and quickly growing amount of images and videos available on the Internet is a very interesting source of data in general and for computer vision research in particular, *cf*. [Goesele10b]. Internet images promise variability on a level that has not been considered almost a decade ago. They can be seen as an—irregularly sampled—approximation to the "space of all images" which makes it well-suited to explore the limits of current algorithms.

Reconstruction algorithms can benefit from this abundance of data. For example, we can assume that almost every interesting part of a tourist landmark is covered by several photos from different distances, angles, and under all kinds of weather conditions. If artifacts such as shadows or occlusion interfere with the reconstruction, we can just discard these images and use others that show the parts in question. The task is then to find and select those images that give the best reconstruction results.

On the other hand, the sheer amount of data and the irregular distribution of images lead to new challenges. In addition, the data is no longer captured by a single person but by multiple observers at different times and with different cameras. Testing reconstruction techniques on this kind of data will hopefully lead to more robust methods for both controlled and uncontrolled capture setups. A lot of known techniques rely on simplified models and assumptions that do not hold for Internet data. These models need to be extended or replaced to reflect the larger space of possible scene appearances. Similarly, the calibration step to obtain the often required prior knowledge has to be adapted.

In the future, Internet data will become increasingly more relevant from an application point of view. Today user-created 3D content already receives a lot of interest as demonstrated by platforms such as Photosynth and Google Earth. Improved image-based reconstruction methods will make this technology even more accessible and common-place.

1.1.2 Challenges

A lot of challenges arise for photometric reconstructions both on the theoretical and on the practical side. On the former, ambiguities inherent to the inverse problem need to be resolved. For instance, a bright surface under a dim light yields the same image as a dark surface under a bright light. One of the challenges on the practical side is that the same scene captured with cameras from different manufacturers usually results in slightly different images, *cf.* [Kim12]. We discuss some of the challenges that are commonly encountered and that are most relevant for this thesis in more detail.

Unknown pixel-correspondences: If several pictures are used for reconstruction, they need to be aligned properly. For a surface point P that is observed in one image at pixel location p, we need to know the corresponding pixel q in another image. Most photometric techniques therefore rely on a capturing setup where all pictures are taken from the same view-point with a fixed camera. These correspondences are then trivially determined as $p = q$. Given pictures taken from several positions, the alignment is a challenging step in itself. If we consider the camera location and viewing direction as given, it is equivalent to knowing the 3D geometry of the object. On the other hand, the geometry—or at least its derivative—is one of the scene components we want to reconstruct.

Non-trivial and unknown illumination conditions: If the lighting consists of a single point light source with known location, photometric stereo works pretty well, as we will show in Section 4.4. Not knowing the light source adds additional degrees of freedom to the inverse problem, making it more challenging and requiring further constraints. More complex lighting scenarios than a single point light source require the integration over all incoming light directions at each surface point, which is difficult to invert. If the illumination is both, complex and unknown, the challenge is even harder.

Non-trivial and unknown reflectance properties: A common assumption in shape and orientation reconstruction is that of Lambertian reflectance. This leads to simple formulations which are, however, only strictly valid for diffuse surfaces and do not apply to a lot of objects, *e.g.* a glossy coffee mug. For reflectance recovery, such an assumption does not make sense and more complex models are used. Recovering the parameters of these models often requires the object shape to be known beforehand, *e.g.* [Lensch01]. Thus, if shape and reflectance are unknown and the material does not comply to the Lambertian assumption, the reconstruction is especially difficult.

Cameras as measurement devices: Apart from a few exceptions, most techniques rely on linear measurements of photometric quantities, *e.g.*, luminance. Consumer cameras are not made to perform highly accurate measurement tasks but to produce visually pleasing images. Their sensor properties and internal post-processing steps are unknown and can introduce a non-linear relationship between the desired quantity and the actually observed data. Neglecting this effect can lead to drastic errors in the reconstructed surfaces as shown by Mongkulmann *et al.* [Mongkulmann11].

Furthermore, the finite amount of photo elements in a sensor dictates a discretization of space, and the analog-digital conversion results in a discretization of the luminance measurements. Finally, limited dynamic range will lead to over- or under-exposed pixels and the loss of detail in not correctly exposed regions.

Outliers and violations of the model: As with any experiment, the collected data may contain erroneous measurements for various reasons, *e.g.* sensor noise or compression artifacts. Similarly, even perfect input data might not fit the assumed model because it is only an approximation of the real world. Certain effects, *e.g.* cast shadows or interreflections, might just not be explainable within the theoretical model. We declare all these cases as outliers. Any photometric technique operating on uncontrolled, imperfect data has to define ways to detect and treat them.

Other challenges exist—especially considering Internet images as input data. Examples are occluders in front of a surface, multiple objects and depth discontinuities, dynamic scene content, *etc.* They will be discussed where appropriate. The ideas presented in this thesis address several of the challenges mentioned, but cannot cover the whole range. Most importantly, we do not consider information on different scale levels that arise when closeups and wide shots of an object are combined.

1.2 Contributions and Overview

We study image-based reconstruction approaches both under laboratory conditions and on Internet data. Such techniques are rarely fully self-contained. They rely on some kind of calibration as additional input. We investigate both components and provide novel ideas to address the challenges that arise in each step. In summary, our main contributions are:

- We discuss how well absolute luminance can be predicted on Internet images based on the Moon as a reference object. Applying the same model, we then predict luminances on general images and show how they can be exploited for perceptual tone mapping of low dynamic range images.
- We present two novel approaches to acquire the position of a point light source if it cannot be assumed far away from the scene. We then evaluate the proposed techniques and the traditional light calibration method with respect to different scene configurations.
- We give a detailed account of the error sources that arise in calibrated photometric stereo. In a series of experiments, we quantify the individual contributions. Based on these findings, we perform an error analysis and draw conclusions about the overall quality of the experimental results.
- We propose the first photometric stereo technique operating on Internet webcam images and non-Lambertian objects. Developing an image selection scheme geared towards photometric reconstructions makes the large amount of data tractable. We present a calibration pipeline for uncontrolled data as a preprocessing step and an outdoor image formation model. The approach yields not only surface orientation but also reflectance properties for the scene.

- Calibrating the camera and light source is important for many photometric techniques but is non-trivial on Internet data. We present a reconstruction approach that does not require calibration apart from the camera position. Instead, we use additional images and a multi-view stereo algorithm to recover a partial scene representation which replaces the reference object in example-based photometric stereo.

- While basic photometric stereo assumes a fixed camera, we show how shading information from changing illumination *and* varying camera positions can be exploited for reconstruction in the lab. Placing a reference object in the scene avoids light calibration and allows handling of non-Lambertian BRDFs.

The structure of this thesis corresponds roughly to the presented contributions. We first give some background information to make the work self-contained. In Chapter 3, we provide an overview over the fundamentals and the state-of-the-art in photometric reconstruction techniques. Chapter 4 considers camera and light calibration mostly under controlled conditions, but also radiometric calibration on Internet images. Parts of this chapter are based on two publications: "How Bright is the Moon? Recovering and Using Absolute Luminance Values from Internet Images" presented at the fourth Computational Color Imaging Workshop 2013 and "Geometric Point Light Source Calibration" presented at VMV 2013. We also include a discussion of the error sources in calibrated photometric stereo.

In Chapter 5, we shift the focus towards webcam data and show how to obtain a calibration based on geo-location and image time stamps. Modeling illumination with a sky model then allows us to recover normals and non-Lambertian reflectance. This chapter extends the publication "Photometric Stereo for Outdoor Webcams" presented at CVPR 2012 and provides additional results and discussion. A simplified version of the image formation model was previously presented in the bachelor thesis by Langguth [Langguth10], which the author supervised. Langguth's work does not contain the crucial image selection part.

We again consider webcams in Chapter 6 where we present a different reconstruction technique based on appearance matching instead of optimizing parametric models. The author of this thesis proposed the general idea as a topic of a master's thesis, which was then implemented for controlled datasets by Ritz [Ritz09]. Later, a considerable extension was published in the paper "Removing the Example from Example-Based Photometric Stereo" at the ECCV workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments in 2010. Chapter 6 contains much more detailed results and shows novel aspects of the approach.

Building on the idea of appearance matching, the technique in Chapter 7 finally allows the camera to be moved between images and reconstructs not only surface orientation but also depth. This method moves away from an explicit calibration pipeline and instead relies on an example object captured under the same lighting conditions. The corresponding paper "Multi-View Photometric Stereo by Example" is currently under review.

We summarize and discuss our work in Chapter 8. Finally, we close with a perspective for future developments.

Chapter 2

Background

Several concepts, notations, and definitions related to scene appearance will recur throughout this thesis. In this chapter, we first recapitulate some information about light and summarize the most important photometric terms. We then introduce the underlying equations for standard photometric stereo techniques. Finally, we define the camera model that we assume and which we will use in later chapters.

2.1 Model of Light

For centuries, scientists have tried to define and explain the nature of light. From a physicist's perspective, the modern definition of light is usually based on its electromagnetic properties. In other disciplines, the physical properties are less important than the more abstract effects that we prescribe to light, *e.g.* perceived brightness, color, tone, *etc.* Here, we will look briefly at both these aspects.

2.1.1 Radiometry

Energy

Whenever energy is transferred or transformed, it is usually important to know the rate at which this happens. For example, a typical light bulb transforms $\Delta W = 360\,\mathrm{kJ}$ of electrical energy in $\Delta T = 1\,\mathrm{h} = 3600\,\mathrm{s}$. The rate $\Delta W/\Delta T$ is called *power*, *e.g.* of a light wave, and describes the temporal change of energy. If we are interested in the spatial component of energy, we define *energy densities* $\Delta W/\Delta V$ or $\Delta W/\Delta A$ for the energy contained in a volume or area respectively. Power and any of these densities can also be interpreted as differentials in the limit of ΔT, ΔV, or ΔA approaching zero.

Electromagnetic Waves

Visible light can be described as electromagnetic waves—or particles—with wavelength λ in the range of about 400 nm to 700 nm. These waves propagate through space, carry energy, and can interact with matter, thereby transferring some of their energy. For most cases relevant in this thesis, we assume a monochromatic, linearly polarized wave in vacuum.

From Maxwell's equations in vacuum, it follows that the electric and magnetic fields fulfill a wave equation

$$c^2\nabla^2 E = \partial_{tt} E, \qquad c^2\nabla^2 B = \partial_{tt} B \tag{2.1}$$

where c is the speed of light in vacuum. A possible solution of these equations are plane waves. For the electric field, we obtain

$$E(x,t) = E_0 \cos(\omega t - \langle k,\ x\rangle + \varphi) \tag{2.2}$$

with angular frequency $\omega = 2\pi c/\lambda$, phase φ, amplitude vector $E_0 \in \mathbb{R}^3$, and a propagation direction $s = \lambda k/2\pi$ where k is called *wave vector*. Again from Maxwell's equations, we find that $E \perp k$, $B \perp E$, and $B \perp k$, *cf.* [Zinth98]. Thus, both the electric and magnetic fields "propagate" along the direction s and are perpendicular to each other. Also, they are in phase.

Energy Transport

The electric and magnetic field can act upon hypothetical charges or currents and thus carry energy. The (volumetric) energy density of a light wave is given as

$$u(x,t) = \frac{1}{2}\epsilon_0 \|E(x,t)\|^2 + \frac{1}{2\mu_0}\|B(x,t)\|^2 \tag{2.3}$$

and follows a wave equation itself. ϵ_0 and μ_0 are physical constants that describe the electric permittivity and magnetic permeability in vacuum. The energy density also propagates along s, which means that

$$u(x,t_0) = u(x + v(t_1 - t_0)\cdot s, t_1) \tag{2.4}$$

for the speed of propagation v.

Imagine a hypothetical rectangle A perpendicular to s and a box $V = v(t_1 - t_0)\cdot A$ created by a small displacement along $-s$. Then, the whole energy

$$W = \int_V u(x,t_0)dx \tag{2.5}$$

is transported through the surface in time $T = t_1 - t_0$ and defines the power per surface area

$$M = \frac{W}{T\cdot A}. \tag{2.6}$$

This quantity is called *irradiance* and can again be interpreted as a differential in the limit of T, A approaching zero.

If the box is small enough for $u(x,t)$ to be constant within V, then the overall energy is

$$W = V \cdot u_{const}. \tag{2.7}$$

Now, assume that the rectangle is slanted in one dimension at an angle θ with respect to s. We create a sheared box by displacements vT along this novel direction $-\tilde{s}$. The volume is then $\tilde{V} = vT\cdot A\cos\theta = V\cos\theta$, and the overall energy is

$$\tilde{W} = \tilde{V}\cdot u_{const} = V\cos\theta\cdot u_{const} = W\cos\theta, \tag{2.8}$$

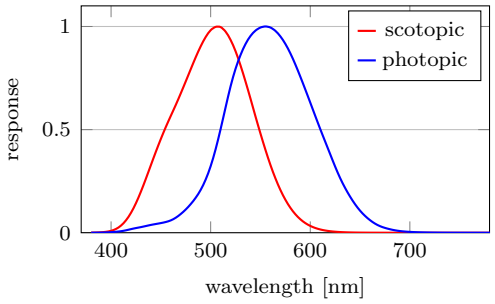


Figure 2.1: Spectral sensitivity of a human observer under scotopic and photopic conditions, *cf*. [UCL].

assuming that the density is constant within $V \cup \tilde{V}$.

If many waves (with $v_i = \text{const}$) pass through the surface from different directions, we obtain

$$W = \sum_i V \cos\theta_i \cdot u_i = \sum_i W_i \cos\theta_i \tag{2.9}$$

where W_i are the respective energies "normalized" to the transfer parallel to the normal of the rectangle. The power per surface area is given as before:

$$M = \frac{W}{T \cdot A} = \sum_i \frac{W_i}{T \cdot A} \cos\theta_i = \sum_i M_i \cos\theta_i. \tag{2.10}$$

In the limit of infinitesimal quantities this leads to the formulation

$$M = \int R(\theta, \phi) \cos\theta d\omega \tag{2.11}$$

where the *radiance* R represents power per area and solid angle.

2.1.2 Photometry

The response of the photosensitive cells in the human eye depends on the wavelength of incoming light. The *cones*, which are responsible for the perception of color, have an average sensitivity as shown in Figure 2.1. It is common to convert the radiometric quantities defined in the previous section into *photometric* quantities that take this sensitivity into account. For example, the radiance is transformed into *luminance* by multiplication with the wavelength dependent sensitivity φ:

$$L(\lambda) = K_m \cdot \varphi(\lambda) R(\lambda) \tag{2.12}$$

where $K_m = 683\,\text{lm/W}$, *cf*. [NIST08].

The cones contribute to our perception mainly for an overall adaption level of $1\,\text{cd/m}^2$ to $10^6\,\text{cd/m}^2$ (*photopic vision*). *Scotopic vision*, *i.e.* low light scenarios at levels of $10^{-6}\,\text{cd/m}^2$ to $10^{-2}\,\text{cd/m}^2$, is dominated by another type of cell. These *rods* have a slightly different sensitivity φ' and thus lead to a different definition of luminance. If not noted otherwise, we always assume the photopic version of all photometric quantities. In the *mesopic* range of $10^{-2}\,\text{cd/m}^2$ to $1\,\text{cd/m}^2$, perception is based on a mixture of rod and cone contributions.

2.1.3 Reflections

When light interacts with matter (on a macroscopic level), it can be absorbed, transmitted, or reflected. Usually all of these effects occur simultaneously, but we will only consider the fraction of light that is reflected.

Let p be a surface point with normal n. The solid angle of an object with respect to a point p is the surface area of its projection onto a unit sphere around p. Let $L(D_{out})$ be the total luminance leaving p into a small solid angle $d\omega_{D_{out}}$ centered around D_{out}. We know that the illuminance incident on a small patch centered at p and perpendicular to n is

$$E = \int_\Omega L_s(D_{in})\langle n,\ D_{in}\rangle d\omega_{D_{in}} =: \int_\Omega t(D_{in}) d\omega_{D_{in}} \tag{2.13}$$

for incoming luminance L_s. Assuming infinitesimally small solid angle and area, we interpret t as the "fraction" of illuminance E caused by incoming light from direction D_{in} or as "angular illuminance density".

Similarly, we define $s(D_{out}, D_{in})$ as the infinitesimal fraction of outgoing luminance caused by reflecting the incoming light from direction D_{in}, *i.e.*

$$L(D_{out}) = \int_\Omega s(D_{out}, D_{in}) d\omega_{D_{in}}. \tag{2.14}$$

According to Nicodemus [Nicodemus77], $s(D_{out}, D_{in})$ is proportional to $t(D_{in})$:

$$L(D_{out}) = \int_\Omega \rho(D_{out}, D_{in}) t(D_{in}) d\omega_{D_{in}} \tag{2.15}$$

$$= \int_\Omega \rho(D_{out}, D_{in}) L_s(D_{in}) \langle n,\ D_{in}\rangle d\omega_{D_{in}}. \tag{2.16}$$

The proportionality factor ρ is called *bidirectional reflectance-distribution function* (BRDF).

2.1.4 Spherical Coordinates

Throughout this thesis, we switch freely between a representation of directions in spherical coordinates (r, θ, ϕ) and Cartesian coordinates (x, y, z) with the transformation defined as

$$x = r \sin\theta \cos\phi, \tag{2.17}$$

$$y = r \sin\theta \sin\phi, \tag{2.18}$$

$$z = r \cos\theta, \tag{2.19}$$

and $r \in [0, \infty[, \theta \in [0, \pi], \phi \in [0, 2\pi[$.

If all directions are measured in a coordinate system with the z-axis equal to n, then Equation (2.16) transforms into

$$L(\theta_{out}, \phi_{out}) = \iint \rho(\theta_{out}, \phi_{out}, \theta_{in}, \phi_{in})\, L_s(\theta_{in}, \phi_{in})\, \cos\theta_{in}\, \sin\theta_{in}\, d\theta_{in}\, d\phi_{in}. \tag{2.20}$$

2.1.5 Simplifications

The reflection formulas simplify for a uniform light source $L_s(D_{in}) = c$ and Lambertian reflectance $\rho(D_{out}, D_{in}) = \alpha/\pi$. The latter also implies that L is independent of the direction D_{out}:

$$L = \frac{\alpha}{\pi}\, c \int_\Omega \langle n,\ D_{in}\rangle \, d\omega_{D_{in}}. \tag{2.21}$$

For an orthogonal matrix R, we can transform the integral into

$$L = \frac{\alpha}{\pi}\, c \int_\Omega \langle Rn,\ RD_{in}\rangle \, d\omega_{D_{in}} \tag{2.22}$$

$$= \frac{\alpha}{\pi}\, c \int_{R^\mathsf{T}\Omega} \langle Rn,\ \tilde{D}_{in}\rangle \, d\omega_{\tilde{D}_{in}} \tag{2.23}$$

because $RR^\mathsf{T} = \mathrm{id}$ and $|\det R| = 1$. This change of coordinates is often useful for R chosen such that $Rn = (0, 0, 1)$, which transforms the global coordinate system into the local one with the z-axis equal to n. However, the resulting integral is still hard to solve for arbitrary sets Ω:

$$L = \frac{\alpha}{\pi}\, c \int_\Omega D_{in,3} \, d\omega_{D_{in}}. \tag{2.24}$$

A further simplification arises if we assume the light source to be an ideal point light modeled as a delta function $L_s(D_{in}) = c \cdot \delta(D_{in} - D_s)$. Then, the integral in Equation (2.16) vanishes:

$$L(D_{out}) = \int_\Omega \rho(D_{out}, D_{in})\, c\, \delta(D_{in} - D_s)\, \langle n,\ D_{in}\rangle \, d\omega_{D_{in}} \tag{2.25}$$

$$= \rho(D_{out}, D_s)\, c\, \langle n,\ D_s\rangle \tag{2.26}$$

and for a Lambertian point it turns into

$$L = \frac{\alpha}{\pi}\, c\, \langle n,\ D_s\rangle. \tag{2.27}$$

2.2 Lambertian Photometric Stereo

Observing a Lambertian surface point M times under distant point light illumination with varying directions $D_{s,1}, \ldots, D_{s,M}$, yields luminance values

$$L = \begin{pmatrix} L_1 \\ \vdots \\ L_M \end{pmatrix} = \begin{pmatrix} \frac{\alpha}{\pi}\, c\, \langle n,\ D_{s,1}\rangle \\ \vdots \\ \frac{\alpha}{\pi}\, c\, \langle n,\ D_{s,M}\rangle \end{pmatrix} = \frac{\alpha}{\pi}\, c \underbrace{\begin{pmatrix} D_{s,1,1} & D_{s,1,2} & D_{s,1,3} \\ & \vdots & \\ D_{s,M,1} & D_{s,M,1} & D_{s,M,2} \end{pmatrix}}_{=:A} \cdot n. \tag{2.28}$$

For $M = 3$ and linearly independent directions $D_{s,i}$, we can invert A and write

$$A^{-1}L = \frac{\alpha c}{\pi}\, n \quad \overset{\|n\|=1}{\Longrightarrow} \quad \|A^{-1}L\| = \frac{\alpha c}{\pi}. \tag{2.29}$$

Thus, n can be computed given the matrix A and the observed luminance L:

$$A^{-1}L = \|A^{-1}L\| \cdot n \quad \Longrightarrow \quad n = \frac{A^{-1}L}{\|A^{-1}L\|}. \tag{2.30}$$

Explicit formulas for the inverse of a 3×3 matrix exist and make a closed form solution possible.

For $M > 3$ and three linearly independent directions, the linear system is overdetermined but has full rank. Thus, $A^\mathsf{T}A$ is invertible, which we exploit by multiplying Equation (2.28) with $(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ (the so called *pseudoinverse*):

$$(A^\mathsf{T}A)^{-1}A^\mathsf{T}L = \frac{\alpha c}{\pi}(A^\mathsf{T}A)^{-1}A^\mathsf{T}A \cdot n = \frac{\alpha c}{\pi}\, n. \tag{2.31}$$

The normal can again be computed from A and L:

$$n = \frac{(A^\mathsf{T}A)^{-1}A^\mathsf{T}L}{\|(A^\mathsf{T}A)^{-1}A^\mathsf{T}L\|}. \tag{2.32}$$

If the incoming luminance c is known, we can also compute the reflectance and vice versa

$$\alpha = \frac{\pi}{c}\|(A^\mathsf{T}A)^{-1}A^\mathsf{T}L\|, \qquad c = \frac{\pi}{\alpha}\|(A^\mathsf{T}A)^{-1}A^\mathsf{T}L\|. \tag{2.33}$$

2.3 Camera Model

We observe the luminance of a scene through a measuring device such as a camera. To abstractly describe the processes involved in converting luminance leaving a scene point into a digitized measurement, we now introduce the camera model that underlies most of the approaches we present later.

2.3.1 Geometric Camera Model

We assume that the physical world can be represented by an affine space defined over the three-dimensional Euclidean space. A digital camera maps three-dimensional points onto a two-dimensional *image plane* which contains the sensor.

Perspective Cameras

We model an ideal pinhole camera simply as a center of projection and an image plane. For now, we assume that the camera is located at the origin and looks along the positive z-axis in a right-handed coordinate system b_1, b_2, b_3. The image plane is perpendicular to b_3 and located at a distance f. It represents a two-dimensional affine subspace, for which we define a coordinate system with basis vectors b_1, b_2 and origin $f \cdot b_3$. Figure 2.2 illustrates this setup from a side view.

The projection of a point $X = (x_1, x_2, x_3) \in \mathbb{R}^3$ onto the image plane is then given by the intersection of its ray through the origin and the plane:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \mapsto \begin{pmatrix} f\frac{x_1}{x_3} \\ f\frac{x_2}{x_3} \\ f \end{pmatrix} =: \begin{pmatrix} u_1 \\ u_2 \\ f \end{pmatrix}. \tag{2.34}$$

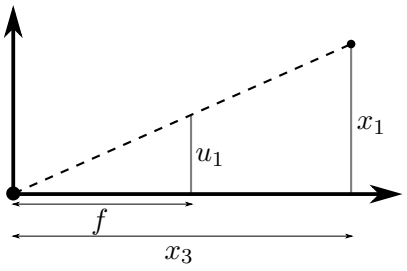


Figure 2.2: The perspective camera model illustrated in 2D. A point X in camera coordinates is projected onto the image plane at f.

The quotient in the image coordinates $(u_1, u_2) = (fx_1/x_3, fx_2/x_3)$ can be made implicit if we associate a vector $(x_1, \ldots, x_n) \in \mathbb{R}^n$ with its equivalence class $\lfloor (x_1, \ldots, x_n, 1) \rfloor$ in projective space $\mathbb{P}^n$ and vice-versa:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f\,\frac{x_1}{x_3} \\ f\,\frac{x_2}{x_3} \end{pmatrix} \leftrightarrow \left\lfloor \begin{pmatrix} f\,\frac{x_1}{x_3} \\ f\,\frac{x_2}{x_3} \\ 1 \end{pmatrix} \right\rfloor = \left\lfloor \begin{pmatrix} f\,x_1 \\ f\,x_2 \\ x_3 \end{pmatrix} \right\rfloor = \left\lfloor \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \right\rfloor . \tag{2.35}$$

The matrix appearing in this formulation is called the *calibration matrix* of the camera.

To model a finite, rectangular sensor, we will use *pixel coordinates* (p_1, p_2) instead of (u_1, u_2). Let the scale factors from image plane coordinates to pixel coordinates be α_1, α_2. The center (c_1, c_2) of the sensor with respect to the origin of the image plane is called *principal point*. Thus, the overall transformation is given as

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \alpha_1\, u_1 + c_1 \\ \alpha_2\, u_2 + c_2 \end{pmatrix} \leftrightarrow \lfloor K \cdot X \rfloor , \qquad K := \begin{pmatrix} \alpha_1 f & 0 & c_1 \\ 0 & \alpha_2 f & c_2 \\ 0 & 0 & 1 \end{pmatrix} . \tag{2.36}$$

We call the offsets c_1, c_2 and the focal length $\alpha_1 f, \alpha_2 f$ *intrinsic parameters* of the camera.

Distortion

The pinhole model is an abstraction, and real cameras or lenses may of course deviate from this concept. For example, real apertures are not infinitely small and thus lead to defocus blur for scene points outside the depth of field. We also do not consider scales where wave effects, *e.g.* diffraction or interference, have to be taken into account and base our discussion solely on geometric optics. These deviations from the model are usually sufficiently small to make the pinhole still an acceptable approximation in most computer vision applications.

Incorporating these effects in the model can, however, become necessary if a closer match to reality is desired. The most commonly considered deviation is radial distortion, which can be quite pronounced for some lenses. This type of distortion projects a point X either closer to the principal point (*pincushion distortion*) than an ideal pinhole camera would do or farther away (*barrel distortion*), depending on the distance of the projection from the principal point. A simple but effective model, *cf.* [Hartley06],

is provided by

$$\begin{pmatrix}\tilde{p}_1\\ \tilde{p}_2\end{pmatrix} = D(r)\cdot\begin{pmatrix}p_1\\ p_2\end{pmatrix}, \qquad r^2 := (p_1-c_1)^2+(p_2-c_2)^2 \tag{2.37}$$

where $\tilde{p}_1, \tilde{p}_2$ are the distorted pixel coordinates and p_1, p_2 are those that would ensue from an ideal perspective projection. When we take radial distortion into account, we will assume that the function D is defined by two parameters k_1, k_2 according to

$$D(r) = 1 + k_1 r^2 + k_2 r^4. \tag{2.38}$$

We will discuss ways to estimate k_1, k_2 and all other camera parameters in Section 4.1.

World to Camera Transformation

Next, we abandon the assumption of the camera being located at the origin. Assuming it is centered at location C with a rotated coordinate system R, we first transform a point $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) \in \mathbb{R}^3$ into the local coordinate system of the camera:

$$X = R\cdot\tilde{X} - R^\mathsf{T}\cdot C. \tag{2.39}$$

The rotation matrix R has three degrees of freedom. Together with the camera center C it represents the six *extrinsic parameters* of the camera.

The overall projection for the perspective camera amounts to

$$\tilde{X} \leftrightarrow \left\lfloor\begin{pmatrix}\tilde{X}\\ 1\end{pmatrix}\right\rfloor \mapsto \left\lfloor K\cdot\left[R\middle|-R^\mathsf{T}C\right]\cdot\begin{pmatrix}\tilde{X}\\ 1\end{pmatrix}\right\rfloor \leftrightarrow \begin{pmatrix}p_1\\ p_2\end{pmatrix}. \tag{2.40}$$

Thus, the camera is completely defined by the *camera matrix* $P := K\left[R\middle|-R^\mathsf{T}C\right]$.

Orthographic Cameras

We have introduced the pinhole as a model for a perspective camera. Other camera models and generalizations, *e.g.* using non-rectangular sensors, exist but will not be of relevance for this thesis. The only exception is the *orthographic camera* which captures parallel rays instead of rays through a center point. Loosely speaking, we can think of a perspective camera that is placed far away from the scene, *i.e.* $\lim x_3 = \infty$, and has a large focal length, *i.e.* $\lim f = \infty$. Then, Equation (2.35) converges to $(u_1, u_2) = (x_1, x_2)$.

More rigorously, we define an orthographic camera as a parallel projection onto the image plane combined with a scaling to obtain pixel coordinates and a possible shift:

$$\begin{pmatrix}p_1\\ p_2\end{pmatrix} = \begin{pmatrix}\alpha_1\, x_1 + c_1\\ \alpha_2\, x_2 + c_2\end{pmatrix} \leftrightarrow \left\lfloor K\cdot\begin{pmatrix}X\\ 1\end{pmatrix}\right\rfloor, \qquad K := \begin{pmatrix}\alpha_1 & 0 & 0 & c_1\\ 0 & \alpha_2 & 0 & c_2\\ 0 & 0 & 0 & 1\end{pmatrix}. \tag{2.41}$$

Again, scene points might need to be transformed into the camera coordinate system before the projection is applied:

$$\begin{pmatrix}p_1\\ p_2\end{pmatrix} \leftrightarrow \left\lfloor K\cdot\begin{pmatrix}R & -R^\mathsf{T}C\\ 0 & 1\end{pmatrix}\cdot\begin{pmatrix}\tilde{X}\\ 1\end{pmatrix}\right\rfloor. \tag{2.42}$$

We note that camera positions which differ only by a translation along the viewing direction will lead to the same projections and cannot be distinguished.

2.3.2 Radiometric Camera Model

The camera transforms continuous luminance values into a digital signal. This transformation includes an integration of radiance over the sensor elements and over time. Furthermore, it encompasses the sensitivity of the sensor, the camera electronics, and possible processing steps such as color space conversions.

We model a camera as consisting of a linear part defined by its optical system and linear sensor, and a non-linear processing part that may contain everything from sensor electronics to post-processing or image editing. The optical system transforms an incoming scene luminance L into the focal plane exposure H which is proportional to the integration time t but decreases with the squared f-number N of the lens. This is expressed as

$$H = \beta \frac{L \cdot t}{N^2} \tag{2.43}$$

with the proportionality factor β. We assume that vignetting effects, which might reduce the luminance at pixels depending on their position on the image plane, can be neglected. These effects usually occur only towards image boundaries.

The signal induced by H is then scaled by the sensor gain ν and processed by the camera to yield non-linear m-bit pixel values:

$$p = (2^m - 1) \cdot f(\nu H) = (2^m - 1) \cdot f\Big(\overbrace{\nu\beta \frac{L \cdot t}{N^2}}^{=:\tilde{p}}\Big). \tag{2.44}$$

We call the monotonically increasing, bijective function $f : [0,1] \to [0,1]$ a *camera response curve*, *cf*. [Mann94, Debevec97].

Relation to Photometric Stereo

If f is the identity function $f(\tilde{p}) = \tilde{p}$, the observed pixel value p depends linearly on the luminance L. Combining Equation (2.28) and Equation (2.44) yields

$$p = (2^m - 1) \cdot \nu\beta \frac{t}{N^2} \cdot \frac{\alpha c}{\pi} A \cdot n =: \tau A \cdot n. \tag{2.45}$$

Similar to Equation (2.30), we can reconstruct the normal

$$n = \frac{A^{-1}p}{\|A^{-1}p\|}. \tag{2.46}$$

Due to the additional unknown factors, however, we cannot recover the reflectance as in Equation (2.33) even if the luminance of the light source was known.

In practice, f is not linear and thus even the normals can no longer be recovered. Ignoring this effect is not an option since it can lead to drastically different results. We illustrate this with a short example for a simplified setting. Let $n = (0, 0, 1)$, $\tau = 1$, and assume that one light shines from above, whereas the other two form an angle $\theta \in]0, \pi/2[$ with the z-axis:

$$A = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & \cos\theta & \sin\theta \\ 0 & 0 & 1 \end{pmatrix}, \qquad A^{-1} = \begin{pmatrix} \frac{1}{\cos\theta} & 0 & -\frac{\sin\theta}{\cos\theta} \\ 0 & \frac{1}{\cos\theta} & -\frac{\sin\theta}{\cos\theta} \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.47}$$

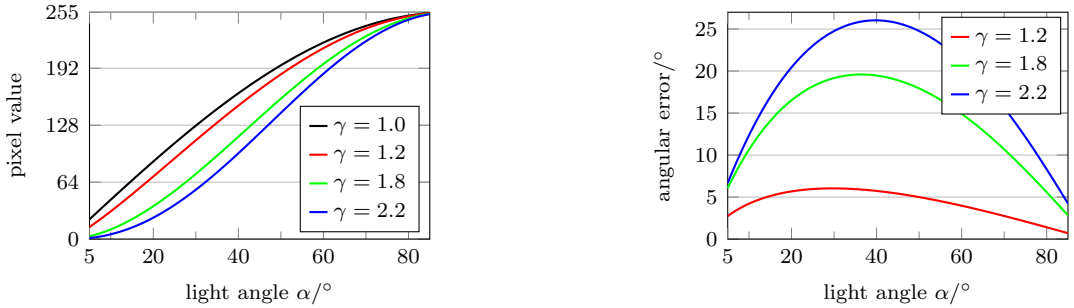


Figure 2.3: Impact of non-linear response on photometric stereo. *Left*: The intensities for the first light direction in a hypothetical setting according to four different gamma parameters. The light direction changes with α. *Right*: The angular deviation of the reconstructed normal if the intensities are not linearized properly.

The ensuing luminance values are $L = (\sin\theta, \sin\theta, 1)$. Applying photometric stereo on the non-linear pixel intensities yields a scaled normal $A^{-1}f(L)$ with norm

$$\|A^{-1}f(L)\| = \sqrt{2 \cdot \Big(\frac{f(\sin\theta) - f(1)\sin\theta}{\cos\theta}\Big)^2 + f(1)^2}. \tag{2.48}$$

Thus, the angular error with the true normal is

$$\operatorname{acos}\Big(\big\langle \frac{A^{-1}f(L)}{\|A^{-1}f(L)\|},\ N\big\rangle\Big) = \operatorname{acos}\Big(\frac{f(L_3)}{\|A^{-1}f(L)\|}\Big) = \operatorname{acos}\Big(\frac{f(1)}{\|A^{-1}f(L)\|}\Big). \tag{2.49}$$

A commonly used—though often oversimplifying—parametric form of the response f is a gamma curve $f(\tilde{p}) = \tilde{p}^{\gamma}$. Figure 2.3 shows $f(\sin\theta)$ and the angular error for varying α. It illustrates the importance of applying the correct inverse response prior to any photometric reconstruction.

Chapter 3

Related Work

Shape and appearance reconstruction have connections to different fields in computer vision, computer graphics, optimization theory, statistics, optics, *etc*. We will only cover those areas in detail that are most relevant for this thesis. These are approaches that rely on varying illumination in multiple images to recover at least the surface orientation and possibly even reflectance and illumination.

That excludes for example the large area of *shape from shading* approaches that operate on single images [Zhang99a, Johnson11, Oxholm12, Han13] and the related works on intrinsic image decomposition [Barron12, Barron13]. These are highly ill-posed problems and their solution requires strong regularization. Using multiple images provides more information and better constrains the result space. We also do not consider purely specular surfaces as in [Healey86, Bonfort03, Chen06, Nehab08, Weinmann13] or the specialized approaches for face capturing [Debevec00, Zhou07, Ghosh11]. We will notice that many reconstruction techniques still rely on laboratory conditions. Those that require complex setups or equipment, *e.g*. [Mueller05, Ma07, Holroyd10], have less potential to be adapted for unconstrained environments. We put a focus on those that have a relatively light-weight capture setup.

Some parts of this thesis are related to reflectance acquisition. This was traditionally achieved through costly measurements of material samples [McNicholas34, Murray-Coleman90]. With the advent of inexpensive digital cameras, it has become possible to speed up this process and to acquire measurements directly from the object of interest [Lensch01, Lensch03, Schwartz11]. These approaches usually require carefully calibrated illumination and known object geometry [Marschner00, Matusik03, Ruiters12]. We limit the discussion to those that jointly recover shape and reflectance.

3.1 Classic Works

At lot of the techniques used today still rely in part on methods and concepts developed 30 years ago. It is instructive to briefly look at some of these early works and give credit to the ideas presented.

One of the first applications of photometry was in exploring the surface of the Moon. Rindfleisch [Rindfleisch65] derives and solves a differential equation for the distance of a surface point from the image plane that depends on the angles of incident and emitted light. His derivation is based on paths in the image plane, and the final

integration depends on at least one depth value along a path to be known. The paths are chosen as straight lines that meet in the point given by the intersection of the image plane with the ray emanating from the camera center in the direction of the Sun. Rindfleisch assumes a certain reflection function for the surface of the Moon that essentially depends on two angles only. Furthermore, the camera geometry must be known and the discussion relies heavily on the Sun as known point light source.

Inspired by that work, Horn [Horn70] formulates the analytical shape from shading problem for arbitrary, but known, isotropic reflectance and known light sources. Again, this amounts to a first order partial differential equation. He transforms this into five ordinary differential equations that are solved along characteristic strips grown from an initial curve. This technique relies on a spatially non-varying reflectance and was only shown to handle relatively simple objects. But the theoretical formulation and the treatment of several special cases that simplify the occurring equations are truly pioneering work. Interestingly, Horn also mentions limitations of his imaging equipment. This aspect has become even more important today as computer vision is employed in consumer hardware with unknown characteristics.

In his PhD thesis—supervised by Horn—Woodham [Woodham77] combines reflectance, illumination, and viewing geometry into a single function, a *reflectance map*, that relates surface orientation directly to image intensities

$$I(x, y) = R(n_{x,y}). \tag{3.1}$$

In general, such a relationship can only be established if each object point receives the same incident illumination, has the same reflectance, and is observed from the same direction. This amounts to a distant light source, an untextured object, and an orthographic camera. The problem of inverting Equation (3.1) in a single image is similar to [Horn70], but instead of formulating a set of differential equations, Woodham defines additional constraints that can guide the inversion process. More important than the single image case is his extension of the photometric shape recovery problem to multiple images under varying illumination. He coins the term *photometric stereo* for a scenario of two or more independent equations

$$I_1(x, y) = R_1(n_{x,y}) \tag{3.2}$$

$$\vdots \tag{3.3}$$

$$I_M(x, y) = R_M(n_{x,y}) \tag{3.4}$$

and explores this concept in the much cited paper [Woodham80]. Since a unit normal $n_{x,y}$ can be described by two angles, Equation (3.2) is an over-determined system of M non-linear equations in two unknowns. In the case of a distant point light source c shining from direction D_i and Lambertian reflectance ρ/π, we obtain

$$R_i(n_{x,y}) = \frac{\rho}{\pi}\, c\, \langle n_{x,y},\ D_i \rangle. \tag{3.5}$$

Inserting into Equation (3.2) yields the formulation in Equation (2.28) which is the basis of many subsequent works in this area.

Horn *et al.* [Horn78] pick up the concept of a reflectance map and give an overview about the possibilities of shape reconstruction from shading in one, two, or three images and compare these with stereo based methods. The authors enhance the

previously mentioned methods to cope with spatially varying albedo by means of ratio images

$$\left.\begin{aligned} I_1 &= \rho(x,y)R_1(n_{x,y}) \\ I_2 &= \rho(x,y)R_2(n_{x,y}) \end{aligned}\right\} \longrightarrow I_{12} := \frac{I_1}{I_2} = \frac{R_1(n_{x,y})}{R_2(n_{x,y})} =: R_{12}(n_{x,y}). \tag{3.6}$$

They also state several properties of stereo and photometric methods that make them complementary techniques:

- Stereo is applied to two or more images from different viewpoints but with constant illumination. Photometric methods work on images from the same viewpoint under varying illumination.
- Stereo methods are suitable for determining the distance of object points whereas photometric methods are more suited when surface orientation has to be recovered.
- Stereo performs best on textured objects with varying albedo whereas photometric methods excel on surfaces with uniform properties.

These points provide the argument for developing hybrid approaches. We will come back to them in Chapter 6 when we demonstrate a possible way of fusing stereo information with photometric data.

In his master's thesis, Silver [Silver80] already notes that the carefully controlled conditions necessary for photometric stereo hinder its broad application:

> "The approach [photometric stereo] is of little value for the unconstrained vision problems facing, for example, a mobile self-supporting robot or biological entity." (W.M. Silver [Silver80])

He develops a look-up scheme that relies on a reference object with the same reflectance as the target. Under varying but otherwise unknown light sources, this allows matching of intensity sequences between target and reference. He transfers the normals from the known reference based on the insight that two surface points with the same normal reflect the same amount of light. We will draw inspiration from this in Chapter 7 where we augment photometric stereo for the multi-view setting.

Coleman and Jain [Coleman82] present a technique that works for non-Lambertian objects without relying on reference objects. They propose to use four known light sources even though three would suffice in the Lambertian case. This allows them to compute a solution for each of the $\binom{4}{3} = 4$ combinations of three light sources. If the surface point is Lambertian, these are very similar and will all lead to the same albedo estimate. If the point exhibits a specularity under one of the sources, these estimates differ and the normal with the lowest albedo is accepted. This procedure relies on the assumption that a surface point behaves almost diffusely under most illuminations and that specularity only arises for a few constellations.

One of the early approaches to consider varying view point and lighting conditions is presented by Hartt and Carlotto [Hartt89]. Their formulation is cast in a probabilistic framework and allows the inclusion of smoothness priors and a model of image noise. Its core is a comparison of observed intensities I with renderings of hypothesized height fields Z. This is already quite similar to many modern approaches. They obtain a solution by Monte-Carlo integration of the posterior $p(Z|I)$ sampled with the Metropolis-Hastings algorithm.

3.2 Uncalibrated Lighting

The first works on photometric methods assumed a known light source. This is a realistic assumption in controlled settings such as an industrial factory or scientific laboratory. Lots of techniques have been developed to provide this kind of input data as we will discuss in Section 4.3. Ultimately, we are interested in uncontrolled settings where this information might not be available.

One of the first methods to cope with this problem was presented by Hayakawa [Hayakawa94] and gave rise to other so called *uncalibrated photometric stereo* approaches. He arranges the luminance at P pixels in M images into a matrix

$$L = \begin{pmatrix} L_{1,1} & \dots & L_{1,P} \\ & \ddots & \\ L_{M,1} & \dots & L_{M,P} \end{pmatrix}. \tag{3.7}$$

Assuming Lambertian reflectance, we know from Equation (2.27) that

$$L = \underbrace{C \cdot D}_{=:T} \cdot \underbrace{N \cdot R}_{=:S} \tag{3.8}$$

where the diagonal matrix $C \in \mathbb{R}^{M \times M}$ contains the source luminance, $D \in \mathbb{R}^{M \times 3}$ represents the light directions, $N \in \mathbb{R}^{3 \times P}$ is the stack of all normals, and $R \in \mathbb{R}^{P \times P}$ contains the reflection coefficients on its diagonal.

For given L, the goal is to find the matrix S. However, multiple candidates $\hat{T}, \hat{S}$ might fulfill

$$L = \hat{T} \cdot \hat{S}. \tag{3.9}$$

In fact, any invertible 3×3 matrix A defines a candidate pair $\hat{T} := T \cdot A, \hat{S} := A^{-1} \cdot S$. Such a candidate pair can be obtained from L using singular value decomposition. Additional constraints are necessary to find the actual S, T.

Hayakawa proposes to use six or more pixels with the same or known albedo, *i.e.*

$$\rho_1^2 = \|s_1\|^2 = \langle s_1,\ s_1 \rangle, \quad \dots, \quad \rho_6^2 = \|s_6\|^2 = \langle s_6,\ s_6 \rangle \tag{3.10}$$

where s_i are the corresponding columns. Choosing the same columns in a candidate matrix $\hat{S}$ yields

$$\rho_i^2 = \|s_i\|^2 = \langle s_i,\ s_i \rangle = \langle A\hat{s}_i,\ A\hat{s}_i \rangle = \langle \hat{s}_i,\ A^\mathsf{T} A \hat{s}_i \rangle. \tag{3.11}$$

These equations constrain the entries of the symmetric matrix $B = A^\mathsf{T} A$, which has six degrees of freedom. Once this system of equations is solved, A can be recovered—up to an unknown rotation—from B using singular value decomposition. Multiplying the candidate with this transformation yields the final result $S = A \cdot \hat{S}$.

Techniques like the one just presented reconstruct a normal field $n(x, y)$ without considering an underlying surface. Instead, Belhumeur *et al.* [Belhumeur99] formulate the problem in terms of a height field, *i.e.* a graph $(x, y, h(x, y))$, with scaled normals

$$n(x, y) = \big(\partial_x h, \partial_y h, -1\big). \tag{3.12}$$

To define a surface, h must satisfy the *integrability constraint*, which had already been used for shape from shading [Horn86b, Frankot88]:

$$\partial_x \partial_y h = \partial_y \partial_x h. \tag{3.13}$$

This provides constraints on any normal field that belongs to a surface. Not all transformed candidates $A \cdot \hat{S}$ are able to fulfill those. Belhumeur *et al.* show that for an integrable normal field S, the set of matrices A that preserve this property is equivalent to

$$\begin{pmatrix} 1 & 0 & -\mu/\lambda \\ 0 & 1 & -\nu/\lambda \\ 0 & 0 & -1/\lambda \end{pmatrix}. \tag{3.14}$$

for parameters $\mu, \nu \in \mathbb{R}$ and $\lambda > 0$. In practice, that means that an integrable normal field can be recovered by photometric methods only up to such a *generalized bas relief transformation* if no additional information is available.

Yuille and Snow [Yuille97] use a similar matrix decomposition approach as Hayakawa but extend the shading model by a constant ambient term

$$L = \begin{pmatrix} L_{1,1} + a_1 & \dots & L_{1,P} + a_P \\ & \ddots & \\ L_{M,1} + a_1 & \dots & L_{M,P} + a_P \end{pmatrix} \tag{3.15}$$

They enforce the integrability constraint to resolve the ambiguity in A up to a generalized bas relief transform. The final solution is then defined by additionally assuming a light source with constant luminance.

These works assumed distant point light sources. Handling arbitrary unknown illumination is usually addressed by a decomposition of incoming luminance. If the incoming illumination L_s can be decomposed into individual components

$$L_s = \sum_{j=0}^{r} \lambda_j L_{s,j}, \tag{3.16}$$

this transfers to the outgoing luminance. Assuming Lambertian reflectance in Equation (2.16), we obtain

$$L = \int_\Omega \frac{\rho}{\pi} \sum_j \lambda_j L_{s,j}(D_{in}) \langle n,\ D_{in} \rangle \, d\omega_{D_{in}} \tag{3.17}$$

$$= \sum_j \lambda_j \underbrace{\int_\Omega \frac{\rho}{\pi} L_{s,j}(D_{in}) \langle n,\ D_{in} \rangle \, d\omega_{D_{in}}}_{=:L_j} \tag{3.18}$$

where L_j is the luminance observed if the scene was illuminated just by $L_{s,j}$. Under these assumptions, the matrix in Equation (3.7) can be decomposed as

$$L = \begin{pmatrix} \lambda_{1,1} & \dots & \lambda_{r,1} \\ & \ddots & \\ \lambda_{1,M} & \dots & \lambda_{r,M} \end{pmatrix} \cdot \begin{pmatrix} L_{s,1}(n_1) & \dots & L_{s,1}(n_P) \\ & \ddots & \\ L_{s,r}(n_1) & \dots & L_{s,r}(n_P) \end{pmatrix}. \tag{3.19}$$

Basri and Jacobs [Basri01a] argue that a decomposition into spherical harmonics represents a good approximation to the space of all possible images of a Lambertian object. A constant illumination yields the zero order harmonic image which corresponds to the surface albedo. The first order harmonic images are taken under cosine lighting for each of the three main axes and correspond to the respective components of the scaled normal at each pixel. Higher order decompositions reduce the approximation error.

In a second paper, Basri and Jacobs [Basri01b] show how to exploit these insights for photometric stereo reconstructions under arbitrary, unknown illumination. They use singular value decomposition to obtain a candidate factorization $L \approx \hat{\Lambda}\hat{L}_s$ as the best rank r approximation of L. As in the discussion about [Hayakawa94], this factorization is only unique up to an $r \times r$ linear transformation. Basri and Jacobs use a normalization constraint to reduce this ambiguity to a Lorentz transformation in the case $r = 4$. Enforcing integrability again leads to a unique solution.

The bas relief transform not only applies to normals but also transforms the diffuse albedo. Alldrin *et al.* [Alldrin07b] exploit the fact that many objects are composed of a small set of albedo values whose histogram gets broadened by a bas relief transform. They define an energy based on the entropy of the albedo distribution. Minimizing this energy yields the parameters of the correct transform.

Favaro and Papadhimitri [Favaro12] look at the function $f(p) := \langle n(p),\ D_s \rangle$ defined over the image domain. They discover that a maximum of this function constrains the parameters μ, ν of the allowable general bas relief transforms to a line and the parameter λ to a semicircle over that line. Intersecting those curves in parameter space for multiple extrema yields a single point which completely describes the transform. Albedo variation can, however, make the detection of maxima difficult.

The works discussed so far assumed an orthographic camera model like the majority of photometric stereo approaches. Papadhimitri and Favaro [Papadhimitri13] show how to incorporate a perspective camera model in a photometric reconstruction. More importantly, they find that a perspective formulation of the uncalibrated photometric stereo problem, when enforcing the integrability constraint, does not suffer from the bas relief ambiguity. Their results indicate also that an incorrect focal length or principal point can lead to strongly biased normals.

None of these approaches considers the radiometric calibration of the camera. It is either assumed that calibration can be performed in a preprocessing step or is circumvented by using special cameras with linear response. Shi *et al.* [Shi10] propose an uncalibrated photometric stereo technique that also recovers the camera response curve. They exploit that the ratio between color channels for a single pixel is constant in a sequence of linear images but curved if a non-linear response is present. Measuring the non-linearity of the profiles in the RGB space allows them to define a minimization problem on the coefficients of a polynomial response model. For photometric stereo, they automatically select pixels with equal albedo and different normals which allows them to remove the bas relief ambiguity as in [Hayakawa94].

3.3 Unknown Reflectance

The approaches in Section 3.2 assume objects of Lambertian reflectance. Most real surfaces have reflectance properties that are neither purely Lambertian nor perfect

mirrors. Several works address this problem of reconstructing shape in the presence of unknown reflectance.

A commonly followed path is to split the luminance into a diffuse and specular contribution

$$L = \hat{\alpha}_{diff} \cdot L_{diff} + \hat{\alpha}_{spec} \cdot L_{spec}. \tag{3.20}$$

Nayar *et al.* [Nayar89b] assume that the specular term consists of a single spike and approximate it with a delta function. That corresponds to a BRDF

$$f(D_{out}, D_{in}) = \hat{\alpha}_{diff} + \hat{\alpha}_{spec} \cdot \frac{\delta(D_R - D_{in})}{\langle n,\ D_{in}\rangle} \tag{3.21}$$

in Equation (2.16) where $D_R = -D_{out} + 2\langle D_{out},\ n\rangle n$ is the reflection vector. They only study the two-dimensional case and reconstruct surface orientation φ_n within the plane formed by D_{in} and D_{out}. Using an extended light source $L_s(\varphi, \varphi_s)$ centered around direction φ_s yields

$$L = \alpha_{diff} \cdot \cos(\varphi_s - \varphi_n) + \alpha_{spec} \cdot L_s(2\varphi_n, \varphi_s) \tag{3.22}$$

where all angles are with respect to the viewing direction D_{out}.

Nayar *et al.* sample this function $L(\varphi_{s,i})$ for multiple known light source directions. At most two samples $i, i+1$ are affected by the specular component because it is a delta peak. From these, α_{spec} and an estimate $\varphi_{n,spec}$ of the normal can be obtained. The others determine α_{diff} and $\varphi_{n,diff}$. They perform this computation for all i and select the result where $\varphi_{n,diff}$ and $\varphi_{n,spec}$ are most similar.

Another way to deal with non-diffuse surfaces is to exploit polarization. For example, Nayar *et al.* [Nayar93] find that in many cases the specular reflection is partially polarized whereas light after diffuse reflection tends to be unpolarized. They can estimate the specular term from multiple images obtained by rotating a polarization filter in front of the camera. Removing its contribution yields an image as if the object was diffuse.

Sato and Ikeutchi [Sato94b] extend [Nayar89b] to a simultaneous analysis of all three color channels in RGB space. Using a similar reflection model, they consider multiple samples in Equation (3.22) and arrive at the matrix expression

$$L = \begin{pmatrix} \cos(\varphi_{s,1} - \varphi_n) & L(2\varphi_n, \varphi_{s,1}) \\ \vdots & \vdots \\ \cos(\varphi_{s,m} - \varphi_n) & L(2\varphi_n, \varphi_{s,m}) \end{pmatrix} \cdot \begin{pmatrix} \alpha_{diff,R} & \alpha_{diff,G} & \alpha_{diff,B} \\ \alpha_{spec,R} & \alpha_{spec,G} & \alpha_{spec,B} \end{pmatrix} =: \mathcal{D}\cdot\mathcal{K}. \tag{3.23}$$

This model corresponds to the *dichromatic reflection model* introduced by Shafer [Shafer84] who notes that the specular vector $(\alpha_{spec,R}, \alpha_{spec,G}, \alpha_{spec,B})$ usually has the same color as the light source. Sato and Ikeuchi estimate this vector from several pixels of different color. Then, the diffuse component $(\alpha_{diff,R}, \alpha_{diff,G}, \alpha_{diff,B})$ is obtained from similar arguments as in [Nayar89b]—most samples capture only the diffuse contribution. Finally, they recover the matrix $\mathcal{D}$ from Equation (3.23) by inverting the color matrix $\mathcal{K}$. To get the actual orientation φ_n, however, knowledge of the light source directions $\varphi_{s,i}$ is required.

Instead of approximating specular reflectance with a delta peak, more sophisticated BRDF models allow for broader lobes and smeared out highlights. Tagare and

deFigueiredo [Tagare91] give theoretical insights on photometric stereo under such "m-lobed reflectance maps" and show experimentally that a simple Lambertian assumption leads to high errors on non-ideal objects. Goldman *et al.* [Goldman05] alternatingly estimate a parametric BRDF model and the surface orientation. Varying the reflectance parameters at each surface point would lead to lots of unknowns, require a huge amount of images, and affect the robustness negatively. Instead, they recover only a small set of basis BRDFs and per-pixel mixing coefficients. The optimization nevertheless requires a good initialization, which they obtain using Lambertian photometric stereo.

Aittala *et al.* [Aittala13] also estimate spatially varying reflectance using a parametric BRDF model—based on sums of Gaussians—and non-linear optimization. In contrast to [Goldman05], they derive one set of parameters at each pixel. Their technique relies on an extended light source, *e.g.* a LCD display, that projects a series of illumination patterns onto the target. It is primarily a BRDF acquisition setup but also recovers surface normals for almost planar objects.

The complexity of the employed BRDF models in these approaches leads to non-linear formulations that are often hard to optimize. Shi *et al.* [Shi12a] introduce a simplified reflectance model that depends linearly on its parameters. They show that it approximates the low-frequency components of other parametric models and measured BRDFs. If the light source positions are known, surface normals and BRDF parameters are reconstructed by alternating two linear least squares optimizations.

Invariants

A different approach to handle complex BRDFs is to exploit general invariants, such as symmetries and physical properties, in the image formation model or capture setup. Such techniques are usually independent of any explicit parametric reflectance model. This is an advantage in terms of generality but can be a disadvantage if editing of scene appearance is desired.

Zickler *et al.* [Zickler02] use *Helmholtz reciprocity*, which states that a BRDF is invariant if the incoming and outgoing directions are swapped: $f(D_{in}, D_{out}) = f(D_{out}, D_{in})$. Their setup consists of a camera and a light source which can be exchanged to acquire a "reciprocal pair". The intensity—ignoring irradiance fall-off due to the light source distance—of a surface point in both images of a pair yields

$$I_r = f(D_{in}, D_{out}) \cdot \langle n,\ D_{in} \rangle, \qquad I_l = f(D_{out}, D_{in}) \cdot \langle n,\ D_{out} \rangle \tag{3.24}$$

$$\implies \quad 0 = \langle n,\ I_r D_{out} - I_l D_{in} \rangle =: \langle n,\ w \rangle. \tag{3.25}$$

For multiple pairs, this leads to the matrix formulation $0 = n^\mathsf{T} W$. Thus, the rank of W is 2 and can be used as an indicator when optimizing for surface depth. Once they find the correct depth, the normal is given by the kernel of W.

Capturing reciprocal pairs requires a careful calibration and camera/light placement. Zickler [Zickler06] shows how such a calibration can be obtained from features in the reciprocal images themselves. It remains, however, necessary to swap the light and camera for each pair. Delaunoy *et al.* [Delaunoy10] use the same concept as [Zickler02] but formulate an energy over a global 3D model instead of separate depth maps.

Alldrin and Kriegman [Alldrin07a] exploit the concept of *bilateral symmetry* for isotropic BRDFs. This requires a circle of light positions—parametrized by the angle

φ—centered about the optical axis of the camera. For a single pixel, the luminance distribution $L(\varphi)$ exhibits a symmetry around a certain angle φ_g. This angle and the viewing direction define a plane that contains the normal. The authors show that this information is sufficient to recover iso-depth contours of the surface. Further constraints are necessary to obtain the unique normal or the absolute depth.

Using this technique as initialization, Alldrin *et al.* [Alldrin08] present a system that recovers not only surface orientation but also BRDFs. This makes it possible to render novel views of captured objects under different illumination. They rely on a set of basis materials which are represented as data samples and thus do not depend on a parametric model. The normals and material weights at each pixel and the basis BRDFs for the full object are found through an alternating optimization.

The bilateral symmetry in [Alldrin07a] determines only the azimuth angle of normals. Shi *et al.* [Shi12b] give an example of an additional constraint to also recover the zenith angle. It is based on an assumed one-dimensional monotonicity of the BRDF and requires light directions to be known and distributed equally on the hemisphere. Holroyd *et al.* [Holroyd08] also investigate symmetries similar to [Alldrin07a], but additionally consider tangents. This makes it possible to study anisotropic materials. They define a measure of symmetry for a pair of (n, t) of normal and tangent under three possible reflections. Minimizing that measure at each pixel yields a local coordinate frame which can, for example, be used to fit a parametric BRDF model. Their technique, however, requires thousands of images with known light position.

Chandraker *et al.* [Chandraker11] again use light sources that move on a circle around the optical axis. They exploit image derivatives both in the spatial and in the temporal domain, *i.e.* neighboring pixels and successive light directions, to define a ratio that is constant over time and independent of the BRDF. From this invariant, it is possible to recover the direction of the gradient and, similar to [Alldrin07a], the iso-depth contours. The advantage of this technique is that it works for unknown light source positions as long as they fulfill the circular motion assumption. A disadvantage in practice is that the reconstruction involves higher-order derivatives of image intensities, which tend to be unstable due to noise. Furthermore, additional input data is needed to completely determine depth and normals.

As in Section 3.2, all techniques discussed here, except [Alldrin07a], rely on pixel values that are linearly related to scene luminance. Higo *et al.* [Higo10] exploit three properties present in many BRDFs which even hold for observations o under an unknown camera response function. From known light source directions D_i, they define a monotonicity constraint

$$o_i > o_j \iff \langle D_i - D_j,\ n\rangle > 0, \tag{3.26}$$

a visibility constraint

$$\langle D_i,\ n\rangle > 0, \tag{3.27}$$

and an isotropy constraint

$$o_1 \simeq \ldots \simeq o_k \quad \Longrightarrow \quad \mathrm{Var}_i\big(\langle n,\ D_i\rangle\big) \text{ minimal.} \tag{3.28}$$

While each of these constraints is relatively weak—visibility is valid for all normals in a half-space—combining them in a consensus manner over lots of light directions yields very good results.

Summary

In this section, we have presented an overview over the most common ways to deal with unknown reflectance: separation into diffuse and specular components, incorporating parametric BRDF models into an optimization, and exploitation of symmetries. The first and second both have the disadvantage of model assumptions that might not be fulfilled by an arbitrary real-world reflectance. They have, however, been shown to work for a broad range of materials in practice. A benefit of parametric BRDF models is that they can be used to create novel impressions of the scene and that they fit into artists' editing pipelines. Approaches based on symmetries only recover shape but are more general because they do not assume a fixed BRDF model. The downside is that they require special constellations of light source and camera during capturing.

A strategy we do not discuss here in depth is to interpret non-Lambertian reflectance as outliers and to apply one of the robust techniques discussed in the next section. This is similar in spirit to the concept of Coleman and Jain [Coleman82] and requires that sufficiently many images show a surface point without specularity.

3.4 Non Ideal Conditions

Several effects, *e.g.* shadows, interreflections, image noise, non-linear camera response, non-ideal light sources, *etc.*, are not accounted for in the simple model employed by most photometric stereo methods. They are often treated as outliers, but sometimes they can even be exploited to obtain additional information. We will discuss some examples of the latter first and then look at outlier treatment.

Nearly all approaches neglect global illumination effects, *i.e.* light that reaches a surface patch not directly from the source but after one or more bounces at other surfaces. The actually observed luminance then has a direct and indirect component $L = L_{direct} + L_{indirect}$ from which only L_{direct} follows the model in Equation (2.27). Nayar *et al.* [Nayar90] discuss the impact of interreflections on photometric reconstruction techniques. They assume the scene to consist of infinitesimal, planar patches of Lambertian reflectance similar to radiosity approaches in computer graphics [Goral84]. The geometric relations between patches can be encoded in a "kernel matrix" K to yield

$$L = (\mathrm{id} - PK)^{-1} L_{direct} = (\mathrm{id} - PK)^{-1} P\, N\, D \tag{3.29}$$

where P, N contain the albedo and normal of each patch and D the light source directions. A standard photometric stereo technique would result in false estimates $\tilde{N}, \tilde{P}$:

$$\tilde{P}\tilde{N} := L \cdot D^{-1} = (\mathrm{id} - PK)^{-1} P\, N \quad \Leftrightarrow \quad P\, N = (\mathrm{id} - PK) \cdot \tilde{P}\, \tilde{N}. \tag{3.30}$$

Nayar *et al.* solve this problem by iteratively updating the left hand side with calculations of the right hand side based on estimates from the previous iteration. Chandraker *et al.* [Chandraker05] study a similar problem in the context of uncalibrated photometric stereo. They are able to show that interreflections actually resolve the generalized bas relief ambiguity.

Shadows are another source of inconsistencies. One distinguishes *cast shadows*—generated because light towards the surface point p is blocked by another point q—and

attached shadows—occurring because the surface faces away from the light source, *i.e.* $\langle n,\ D\rangle \leq 0$. Daum and Dudek [Daum98] reconstruct surface height based on cast shadows from known point light sources. They derive several constraints based on geometric reasoning such as "if point p is shadowed by q, then all surface points on the connecting line must lie below the ray from p to q". Kriegman and Belhumeur [Kriegman01] instead study attached shadows. They reconstruct the light positions and normals at intersection points of shadow boundaries. Okabe *et al.* [Okabe09] exploit the binary pattern (shadowed/non-shadowed) at a pixel created by a large sequence of light positions. The similarity between two such patterns is related to the angular difference of the corresponding normals. This insight allows them to recover surface orientation using dimensionality reduction inspired by [Sato07]. Sunkavalli *et al.* [Sunkavalli10] treat both types of shadows. They observe that the light source visibility is constant in subregions of the surface. The intensities in each region lie in a three-dimensional subspace as exploited by uncalibrated photometric stereo methods. The overall intensity matrix thus consists of several rank three submatrices which can be obtained using subspace clustering techniques.

In contrast to these approaches, another line of work tries to detect any deviation from the local, Lambertian shading model and then applies photometric stereo only to the remaining intensities. This usually requires many images to ensure that a surface point is observed sufficiently often without shadows or specular behavior. Such a "dense photometric stereo" technique is proposed by Wu and Tang [Wu06]. They formulate a probabilistic image formation model that includes a binary inlier/outlier map as hidden variable. For a single pixel in M images, they compute a set of $\binom{M}{M-1}$ normals and represent their distribution as a covariance matrix K. The algorithm estimates an optimal K by weighting down observations that are likely to be outliers according to the map. After updating the weights, the whole process is repeated until convergence. Another example of this category is given by Verbiest and Van Gool [Verbiest08] who also employ expectation maximization but treat the normal map as a hidden variable. Instead of the per-pixel approach of Wu and Tang, they describe the likelihood of outlier intensities with per-image histograms. Beljan, Ackermann, and Goesele [Beljan12] also consider sets of normals for a single point computed from $\binom{M}{3}$ subsets of all available images, but do so in a multi-view setting. For each element, they compute the set of inlier images whose observations are consistent with the hypothesized normal. Instead of studying the distribution of these normal hypotheses, they apply RANSAC to find the one that has the largest support. The number of inliers then provides a cue to decide whether a voxel belongs to the true surface or not. This leads to good results for the normals but turns out to be not very discriminative in terms of absolute geometry.

A slightly different perspective on photometric stereo with outliers is provided by techniques that analyze the observation matrix for errors and missing entries. Wu *et al.* [Wu10] interpret the problem as a low rank matrix recovery task in the presence of sparse corruptions. More formally, they augment the matrix decomposition in Equation (3.8) with an error matrix E that contains all outliers due to shadows or specularities:

$$L = T \cdot S + E =: A + E. \tag{3.31}$$

As we have seen before, the rank of $T \cdot S$ is at most three for a Lambertian sur-

face. Under the assumption that outliers occur infrequently, E is sparse and the task becomes

$$\arg\min_{A,E}\big(\operatorname{rank}(A) + \alpha\|E\|_0\big), \quad \text{s.t. } L = A + E \tag{3.32}$$

where $\|\cdot\|_0$ is the number of non-zero entries—and thus not a true norm. Wu *et al.* replace this formulation with a convex optimization problem

$$\arg\min_{A,E}\big(\|A\|_\star + \alpha\|E\|_1\big), \quad \text{s.t. } L = A + E, \tag{3.33}$$

which can be minimized iteratively using Lagrange multipliers with additional penalty terms ("Augmented Lagrange Multiplier method"). If the locations of shadows are known beforehand, they enforce the constraints only on the remaining entries. Once A is recovered, standard tools such as singular value decomposition can be applied to obtain the factorization S, T. Other approaches in this field of research, *e.g.* [Okatani07, Okatani11, Eriksson10], directly optimize for the decomposition

$$\arg\min_{S,T} \|W \odot \big(L - T \cdot S\big)\| \tag{3.34}$$

with W encoding missing entries and $\odot$ the component-wise multiplication. Such a technique can also be applied to cope with outliers in calibrated photometric stereo as, for example, shown by Ikehata *et al.* [Ikehata12].

3.5 Unknown Lighting and Reflectance

Shape reconstruction of Lambertian surfaces under unknown illumination is possible as discussed in Section 3.2. Techniques that handle more complex BRDFs are presented in Section 3.3 but require known light sources or special capture setups. Reconstructing objects with complex BRDFs under unknown illumination is a much harder problem because both constituents interact to form the observed appearance.

Low-rank factorizations of the intensity matrix as in uncalibrated photometric stereo are no longer easily possible. A straightforward way to deal with this problem is to separate the diffuse and specular contributions using a technique such as [Tan05]. Applying one of the methods in Section 3.2 to the diffuse component then yields the desired normal field. Tan *et al.* [Tan07] propose such a system and show that the generalized bas relief ambiguity can be resolved from the additional information present in the specular component. Resolving the ambiguity requires two images, but Tan *et al.* [Tan09, Tan11] show how this can even be reduced to one. Wu and Ping [Wu13] use a similar strategy of diffuse-specular separation and find the optimal bas relief transform through a constraint on the structure of the BRDF. This requires objects with sufficient distribution of normal directions, *e.g.* a sphere, and an isotropic BRDF that exhibits a symmetry around the halfway vector.

Instead of factorizing a matrix, it is also possible to exploit geometric constraints if present in the capture setup. Lu and Little [Lu95] present a solution in a rather special setting: the light source and camera are co-located and the object spins around a known axis. They derive surface orientations pointing towards the light from intensity maxima and then exploit the known rotation angle of the object to track them. The

tracked points allow the reconstruction of a one-dimensional slice of the BRDF. Once this is known, the surface orientation of all other surface points can be computed.

A different approach is to define a mathematical model of image formation that encompasses all unknowns and to directly minimize a suitable error. This is similar in spirit to [Goldman05] and comes at the same costs: complex optimization problem, susceptible to local optima, and dependence on a parametric model that might not represent all real-world scenarios. Georghiades [Georghiades03] uses a simplified Torrance-Sparrow BRDF [Torrance67] and a single distant light source. He is able to show that the generalized bas relief ambiguity is resolved for specular surfaces in almost all cases if at least four images are captured. The optimization is, however, restricted to a single BRDF and only allows the diffuse albedo to vary spatially.

We have seen another way to approach the problem in Section 3.3: symmetries and general properties of BRDFs. In this setting, reasoning about symmetries suffers from unconstrained lighting. A much more common strategy is to exploit general relations between *observation vectors* or *appearance profiles* that encode the temporal appearance variation of a single pixel, *i.e.* columns in Equation (3.7).

A recurring concept in this context is *orientation consistency* which states that points with similar normals have similar appearance. Assuming an orthographic camera, *i.e.* constant $D_{out} = v$ for all surface points, the luminance of a point P with normal n is given by

$$f^{-1}(I) = L = \int \rho(v, D_{in})\, L_s(D_{in}) \langle n,\ D_{in} \rangle d\omega_{D_{in}} \tag{3.35}$$

according to Equation (2.16). The right hand side depends only on the normal and BRDF but not on the 3D position. Thus, for a point Q on the surface with the same normal and BRDF as P the luminance is the same. Moreover, if the luminance is the same, the pixel intensities are equal, too—even for a non-linear response curve f.

Hertzmann and Seitz [Hertzmann03, Hertzmann05] place one or more example objects of known geometry in the scene to obtain reference profiles of "basis BRDFs". For a pixel on the target object, they match its appearance profile against linear combinations of these reference profiles to find a corresponding point on the reference, usually a sphere. Based on orientation consistency, both points are then likely to have the same normal. Since normals for the reference object are known, they can simply be transferred to the target. This approach has the added advantage to require only a minimal amount of calibration.

Koppal and Narasimhan [Koppal06] obtain the extrema of appearance profiles from their derivatives under unknown but smoothly varying light source positions. They show that most of these extrema are due to constant constellations of viewing direction, light source, and surface normal. This insight makes the set of extrema locations well-suited features for clustering of surface orientations independent of the BRDF. No relationship between iso-normal clusters, *e.g.* one facing left and one facing right, can be recovered. Other techniques are needed to assign absolute orientation to the clusters. Those can, however, benefit from the initial clustering as prior knowledge.

Sato *et al.* [Sato07] look at appearance profiles as a whole and not just at the locations of extrema. They argue that similar profiles—seen as vectors in M-dimensional space—correspond to similar normals, which they justify based on the Torrance-Sparrow BRDF model [Torrance67]. Since normals are defined on the unit sphere

in $\mathbb{R}^3$, this implies that the set of profiles $\mathcal{M}$ actually is a two-dimensional manifold in $\mathbb{R}^M$. Thus, there should exist an embedding $\Phi : \mathbb{R}^M \to \mathbb{R}^3$ with $\Phi(\mathcal{M}) \subset \mathcal{S}$ that preserves the intrinsic structure of $\mathcal{M}$, *i.e.* points that are "close" in terms of $\mathcal{M}$ correspond to normals that are close on $\mathcal{S}$. The notion of "closeness" on $\mathcal{M}$ is expressed by the geodesic distance, which is itself approximated by euclidean distances along a sequence of neighboring profiles.

Sato *et al.* [Sato07] then employ the Isomap [Tenenbaum00] non-linear dimensionality reduction technique to obtain the embedding and afterwards force the result vectors to have unit length. The normal map thus recovered contains only relative orientations, *e.g.* rotation and reflection are not considered, but can be made unique if additional constraints, such as contours, are applied. The advantages of this technique are its generality and a certain robustness against non-linear camera responses. It makes, however, some approximating assumptions about spatially varying reflectance and requires hundreds of images.

Lu *et al.* [Lu13] extend [Sato07] in several ways. They show empirically—using the MERL [Matusik03] database of measured BRDFs—that the geodesic distance of the profiles is proportional to the angular difference of normals and that the proportionality factor depends only on the material. They also propose how this factor can be estimated from intensity samples obtained under uniformly distributed, but unknown, point light sources. Thus, from intensity profiles, it is possible to recover the $P \times P$ matrix $N^\mathsf{T}N$ containing the dot products of all pairs of normals. Instead of dimensionality reduction techniques, their recovery step is based on factorizing this matrix and removing the arising ambiguity by means of the integrability constraint.

Approaches that reconstruct shape from objects with complex BRDF and unknown lighting are less common than the more restrictive settings we have discussed before. This section has presented several examples demonstrating three high-level strategies: diffuse-specular separation, parametric modeling, and relations between appearance profiles. We will make use of concepts from the last category in Chapter 6 and Chapter 7.

3.6 Multi-View Settings

Most photometric techniques assume a fixed view point. Information about the orientation and reflectance of a surface is recovered from changes in the illumination. This, however, restricts the result to a single perspective and does not allow a reconstruction of a full, *e.g.* 360°, object model. Furthermore, varying view points provide additional information. For example, observing an object from the front and side lets us estimate its absolute position in the scene and not just the surface orientation. Multiple views also introduce redundancy in surface coverage, which can be exploited to increase robustness.

Traditionally, reconstructions from two or more views have been addressed by stereo or multi-view stereo methods. The idea is to find properties that are invariant—or almost invariant—to view-point changes such as the color of a Lambertian surface point P under constant illumination. If P projects to pixel p_1 in the first image and p_2 in the second, their color will be the same. For a hypothesized point $\tilde{P}$ that does not lie on the surface, the color of its projections $\tilde{p}_1$ and $\tilde{p}_2$ will differ in most of the cases. Thus, it is possible to detect the correct position by checking the consistency

of projections with respect to the invariant.

Seitz *et al.* [Seitz06] give a good introduction to the aspects of different multi-view stereo approaches and provide an overview over the literature in this field. We note that almost all techniques assume a Lambertian surface. Exceptions are [Jin03, Soatto03, Jin05], which use a rank constraint on the radiance tensor to handle specular reflectance, or the work by Yu *et al.* [Yu04], who compare input images with renderings of an object under a simplified Torrance-Sparrow [Torrance67] model. These approaches still require the illumination to be constant. Some of the first steps towards multi-view reconstructions that consider shading are also connected to the analysis of densely sampled video sequences. Carceroni and Kutulakos [Carceroni01] use multiple video streams at calibrated positions to recover the motion, shape, and Phong reflectance parameters [Phong75] of deformable surfaces under known point light sources. To optimize their high-dimensional image formation model with respect to the observed data, they split the problem into several subcomponents and initialize them through an explicit sampling of the parameter space. Simakov *et al.* [Simakov03] assume that the object motion is known and base their consistency measure on the residual error of the Lambertian model presented in [Basri01b]. Chandraker *et al.* [Chandraker13] derive a general theory for shape reconstruction from differential object motion. It extends optical flow to objects of unknown BRDF under arbitrary illumination. These works require, however, fixed lighting conditions. We will now turn to techniques that actually exploit the changes in illumination even in a multi-view setting.

Zhang *et al.* [Zhang03] introduce a framework that encompasses the similarity of patches seen from multiple cameras and their appearance change caused by varying illumination. The authors recover the direction of a distant light source, the surface normals, and absolute depth—*i.e.* the distance from a camera center—in an alternating optimization. Given the illumination, they compute normals by inverting a Lambertian shading model and turn them into a depth map through integration. This map is adjusted to pass through the positions of tracked feature points.

Lim *et al.* [Lim05] use a similar pipeline of integrating normals into a depth map and then making adjustments. In their case, the normals are obtained from an uncalibrated photometric stereo step, and the positions of tracked scene points allow a disambiguation of the general bas relief transform. The obtained surface is then projected into all views to create a new intensity matrix, which can again be factorized. This process is iterated until convergence. It is initialized with a piece-wise planar surface triangulated from the tracking points.

Joshi and Kriegman [Joshi07] follow the same line of thought. They derive a consistency measure for Lambertian surfaces which allows them to estimate a coarse depth map. Projecting this depth map into all images then yields an intensity matrix, which can be factorized. They resolve the ambiguities by comparing the pseudo normals with the normals obtained by differentiating the depth map. Lastly, the normals and depth map are both integrated into a final surface.

Their consistency measure is based on hypothesizing planar patches, which they project into all images and form an intensity matrix for each patch. As we have seen in Section 3.2, this matrix has rank three for a Lambertian surface. If the patch was at an incorrect depth, the error of a rank three approximation to this matrix should be high. Minimizing this error combined with a straightforward smoothing does not

yield high quality depth maps, but is sufficient as initialization.

We observe that all of these techniques alternate between a depth estimation and a normal estimation step. This is because photometric stereo requires known pixel correspondences to associate the intensities to a surface point. In traditional photometric techniques, these are trivially defined by the fixed view point assumption.

The approaches discussed above employ a local surface representation, *i.e.* defined per view. We will now look at methods that rely on a global geometric model of the surface. This has the advantage of readily available occlusion information while the surface is deformed by the optimization.

Weber *et al.* [Weber02] use a voxel representation and consider objects on a turntable. They carve away voxels for which the predictions of the Lambertian model—given light source positions—disagree with the recorded intensities even for the best fitting normal. The estimated albedo and normals can then be used to render the scene with novel lighting.

Yang *et al.* [Yang03] develop a consistency measure based on the behavior in color space of a surface point observed from multiple views: it varies linearly from the diffuse color to the color of the light source. They also recognize the need for a smoothness constraint. This is not straightforward to define for a voxel representation, and they resort to extending the disparity gradient [Burt80], which is originally defined between two images only. Their method does not recover normals and mainly considers a setting with fixed illumination.

Treuille *et al.* [Treuille04] also reconstruct a voxel representation. They consider multiple light directions and require one or more example objects in the scene as in [Hertzmann03]. Their consistency measure compares appearance profiles for a candidate voxel against those on the reference objects. The reconstructed normals contain much more details than the coarse voxel approximation. This information can, however, only be exploited during rendering because normals and voxels are not fused into a single surface representation. A further restriction concerns the camera positions, which need to be separated from the scene by a hyperplane to ensure the correct order during voxel processing.

Vogiatzis *et al.* [Vogiatzis06] and Hernandez *et al.* [Hernandez08] represent the global model as a triangle mesh with attached normals, which circumvents the drawback of [Treuille04]. The approach in both works assumes Lambertian surfaces and treats specular reflections as outliers. It relies heavily on object silhouettes to recover the camera parameters, light positions, and an initialization of the mesh. The mesh is then projected into the images to obtain the intensities needed to compute normals with a standard photometric stereo approach. This step is alternated with a vertex displacement that forces the mesh normals to conform with the reconstructed ones. Birkbeck *et al.* [Birkbeck06] also filter out specular reflections during shape recovery but fit a specular Blinn-Phong model [Blinn77] once the final geometry is available. They use a very controlled capture system to obtain the input images and the required light calibration. All three techniques can draw on initializations that are already quite close to the true surface.

Yoshiyasu and Yamazaki [Yoshiyasu11] show that a similar approach can succeed with a much simpler initialization. They also operate on mesh vertices, but convert them into an implicit surface and back to a mesh during the optimization. This allows to handle a greater variability in topology—implicit surfaces can, for example,

self-intersect. Similar to [Hernandez08], their technique requires silhouettes to be extracted. In fact, if the mesh deformation is constrained solely by silhouettes—ignoring the photometric term completely—it already provides very good results. The photometric cue is much weaker and only serves to add some details.

Combinations of silhouette cues and shading are also used in dynamic performance capture systems [Ahmed08, Vlasic09, Wu12] where illumination can be controlled. Wu *et al.* [Wu11] instead reconstruct an initial mesh using multi-view stereo and refine it with normals from uncalibrated photometric stereo. Paterson *et al.* [Paterson05] assume an almost planar target which allows them to warp the images in order to establish pixel correspondences. They use a calibrated flash light as basis for the recovery of normals and the parameters of a modified Torrance-Sparrow model. Ruiters *et al.* [Ruiters09] also assume a planar target. They represent geometry variations as a height field and recover spatially varying BRDFs similar to the approach by Goldman *et al.* [Goldman05]—but in a multi-view setting. This approach even incorporates interreflection effects in the micro-structure and is related to BTF ("Bidirectional Texture Function") acquisition techniques.

A disadvantage of approaches that project vertices into images is that the mesh resolution is not necessarily related to image resolution. Either a huge number of vertices has to be used or details in the images might be lost. Park *et al.* [Park13] therefore propose a pipeline that uses mesh parametrization to define warpings from images onto a planar mesh representation. Thus, fine detail—in the form of a displacement map—can be recovered even with few mesh faces. The base mesh is obtained by merging multi-view stereo reconstructions from all views. Again, specular reflectance is treated as an outlier and the underlying photometric stereo technique is [Hayakawa94].

These multi-view photometric stereo approaches can handle changes of both camera and light position in each image—even if Park *et al.* [Park13] use multiple images per camera in their experiments. Zhou *et al.* [Zhou13] require a whole photometric image series for each camera position. This allows them to cope with arbitrary isotropic BRDFs by applying [Alldrin07a] in each view. The resulting iso-depth contours are then associated with sparse structure from motion points to obtain absolute depth values. The authors propagate depth along contours from each view to merge them into a globally consistent model. Once the geometry is acquired, a set of basis BRDFs and their mixing weights can be estimated because illumination is known from calibration spheres.

Tunwattanapong *et al.* [Tunwattanapong13] create a lighting sequence in each view that is shaped as spherical harmonics. This allows them to recover reflectance parameters in addition to surface orientation but needs a complex capture setup. The results from each camera position are fused with a multi-view stereo reconstruction in a final surface optimization, which again alternates between several stages.

Another technique that relies on a special capture setup to acquire series of images in each of multiple cameras is presented by Schuster [Schuster10]. He proposes a photo-consistency measure based on the fitting error of a Cook Torrance BRDF at each hypothetical voxel and extracts a surface using graph cuts [Boykov03, Lempitsky07]. As initialization, he constructs the visual hull to obtain visibility information and computes normals based on Helmholtz stereopsis. This requires the light and camera to swap positions, which is approximated in this case by a hemispherical dome of

151 cameras equipped with flash lights. A similar setup, but augmented with several projectors for structured light acquisition, is used by Weinmann *et al.* [Weinmann12]. They acquire multi-camera, multi-light image sequences of an object rotating on a turntable. The surface is then extracted from an error function that combines "multi-view structured light consistency" and Helmholtz stereopsis into a single variational formulation.

A very good example of heavily model-based approaches is provided by Yoon *et al.* [Yoon10]. Their generative model consists of a finite number of distant light sources and shadow maps, the surface represented as a level set, reflectance parameters for the Blinn-Phong model [Blinn77], and a set of pinhole cameras. Visibility is handled by projecting the global surface into each image. Furthermore, their technique requires a model of the background scene to prevent the surface from shrinking to an empty set during the optimization. They also introduce an auxiliary normal field to decouple the appearance computation from the surface gradient for increased stability. The final optimization minimizes the error of all images compared to the renderings from the current parameters. It has to be performed in an alternating fashion—like in most other works—because of its complexity and the coupling of shape and reflectance in the image formation.

This section demonstrates that, in the multi-view setting, normals are exploited mostly to provide detail information on top of a proxy geometry. They can also help to guide the reconstruction in textureless regions where stereo methods perform poorly. The predominant assumption is, however, a Lambertian BRDF. In Chapter 7, we will present a multi-view technique that works on non-Lambertian reflectance also and removes several restrictions from the method by Treuille *et al.* [Treuille04].

3.7 Internet and Outdoor Images

The sections above show that many of the state of the art approaches draw on ideas and methods that have been developed ten or twenty years ago, *e.g.* matrix factorization or reflectance symmetries. We observe that today, more focus is put on integrating these ideas into ready to use systems and increasing their applicability through combination with other techniques. Two important aspects are the robustness with regard to non-ideal input images and the removal of requirements imposed on the capture setup. Thus, images downloaded from the Internet provide an interesting testbed for new developments. We emphasize this relevance by discussing some of the advances in computer vision related to Internet data.

Snavely *et al.* [Snavely06] show that it is possible to apply robust structure from motion to images from online photo sharing sites, *e.g.* Flickr. They recover the geometric camera parameters and a sparse set of 3D feature points. This allows a user to explore the scene by "3D browsing" of the images and provides a novel way to communicate the impressions of one or several observers. Tompkin *et al.* [Tompkin12] extend 3D browsing to video sequences. Since known camera parameters are a prerequisite for many reconstruction approaches, Snavely *et al.* also paved the way for works in that area. Goesele *et al.* [Goesele07] present a multi-view stereo approach that relies heavily on the selection of suitable images to achieve robustness against scale differences or occluding clutter. Furukawa and Ponce [Furukawa10b] achieve robustness through several filter steps that reason about visibility and consistency of 3D

patches. These approaches recover the full scene geometry instead of a sparse feature set and can convey the 3D impression of the scene even better. Goesele, Ackermann *et al.* [Goesele10a] exploit this as proxy geometry in an image-based rendering system to achieve a more convincing browsing experience in the presence of unreliable data. Agarwal *et al.* [Agarwal09] and Frahm *et al.* [Frahm10] extend the reconstruction approaches to whole city areas using clustering techniques.

A more complete separation of appearance for uncontrolled scenes is achieved by Haber *et al.* [Haber09]. They build on a 3D model as produced by [Goesele07] and recover the reflectance—encoded as mixing weights per surface point with respect to a selection of Cook-Torrance BRDFs—and illumination in each view of an image collection. The optimization alternates between both unknowns and yields a decomposition that gives plausible relighting results. Diaz and Sturm [Diaz11a] also use a 3D model to estimate illumination, but additionally consider the camera response function. They represent this function as a linear combination of the EMOR basis ("Empirical Model of Response" [Grossberg03, Grossberg04]) and include the coefficients in the lighting optimization. While this work recovers only the diffuse albedo, it is one of the few that actually take non-linear response curves beyond a straightforward gamma correction into account. Garg *et al.* [Garg09] extend results from Belhumeur and Kriegman [Belhumeur98] about the space of possible images for a given scene to photo collections. They project all images onto a 3D model and factorize the matrix that arises from stacking intensities at each vertex. The factorization can be interpreted as "basis images" for the scene that capture variation along different axes. While these axes often correspond to a certain meaning, *e.g.* mean image and shading variations in the Lambertian case, there are no clear semantics for decompositions in more complex scenes.

Another growing source of Internet data are video sequences, *e.g.* from webcams. Jacobs *et al.* [Jacobs07a] collect images from hundreds of webcams and discover that the coordinates of a principal component analysis (PCA) for different cameras behave similarly over time. This allows them to define a canonical basis and assign pixel labels with a common meaning in all cameras. In subsequent works, Jacobs *et al.* [Jacobs07b, Jacobs10, Jacobs13a, Jacobs13b] are able to estimate the location of an unknown camera by comparing PCA coordinates with those of known webcams. The depth of individual scene points can be recovered from correlations induced by cloud shadows. Additionally, the movement of the clouds themselves yields cues to predict vanishing points and thus calibrate the camera geometrically. Lalonde *et al.* [Lalonde08, Lalonde10] demonstrate how the sun and sky illumination can also be exploited for such a calibration, see Section 5.4.4.

Apart from calibration and geo-location, webcams are also used to decompose scene appearance and enable editing of its components. Sunkavalli *et al.* [Sunkavalli07] analyze the intensity profiles of pixels in dense video to identify transitions from shadow to sunlight. They find that profiles within a scene are similar up to scale and a shift in time. This allows them to define time-varying basis curves for the sun and sky contribution. The corresponding coefficients with respect to this basis are recovered similar to [Lawrence06] and constitute a decomposition into ambient light, albedo, and sun reflectance. Sunkavalli *et al.* [Sunkavalli08] additionally recover partial surface normals—projected on the plane of solar movement—based on color changes. Lalonde *et al.* [Lalonde09] estimate the sky appearance and sun visibility in

each frame for a collection of webcams. These cues are used to match the illumination conditions between different cameras and thus enable the transfer of correctly lit objects. The authors also insert synthethic objects by illuminating them according to the parameters of the underlying sky model in each image.

Only few approaches focus on surface orientation as an important part of scene appearance. Shen and Tan [Shen09] use a decomposition of illumination as proposed by Basri and Jacobs [Basri01b] to estimate weather conditions in Internet photo collections. This also involves computation of surface normals, but only at sparse feature points that have been matched between images. Their approach relies on a selection of suitable features with sufficiently Lambertian behavior and support in close-by views. Abrams *et al.* [Abrams12] recover dense surface normals, albedo, and the radiometric camera calibration from outdoor webcams. They exploit the known sun position with respect to a geo-referenced coordinate system in order to arrive at a traditional photometric stereo problem for Lambertian surfaces. This problem becomes more complex through an ambient term and the non-linear response, which is represented as a combination of EMOR [Grossberg03, Grossberg04] basis curves. Abrams *et al.* split the corresponding optimization into two alternating steps and approximate each of them with a linear subproblem which can be solved efficiently. This allows them to use hundreds of images and thus increase robustness to outliers that are not captured by their image formation model.

Compared to traditional approaches that require laboratory conditions, Internet images are at least two levels more challenging. It makes sense to also think about in-between settings, *e.g.* a controlled outdoor dataset, which are not yet solved entirely. Sato and Ikeutchi [Sato94a, Sato95] apply their separation of diffuse and specular components based on illuminant color, presented in [Sato94b], to outdoor images under clear skies. They normalize image regions to simulate a uniform albedo and obtain its value from the brightest pixels in the image sequence. Then, the angle between the normal and the solar plane is obtained at each pixel simply as its maximal intensity divided by the albedo. Narasimhan *et al.* [Narasimhan02] study appearance variation due to weather changes in long sequences of controlled outdoor data. Atmospheric scattering which depends on the distance of the scene point from the camera allows them to compute approximate depth for each pixel. Yu *et al.* [Yu13] capture the environmental illumination from a mirror sphere and discretize it into a set of point light sources. They assume a Lambertian reflectance, treat specularities as outliers, and employ a simple heuristic to estimate self occlusion. Their results on outdoor images expose problems due to insufficient variation of light directions. This might be one of the reasons why most current work on outdoor reconstructions is focused on fixed illumination settings. Especially techniques that decompose the illumination and reflectance in images of a known shape show some promising advances. While this thesis focuses more on shape recovery, it also relates to appearance reconstruction, and we will discuss some of these approaches.

If the whole light field of the scene is known, it can be decomposed into a basis of spherical harmonics. Ramamoorthi and Hanrahan [Ramamoorthi01] show that those frequencies of illumination and reflectance that are also present in the outgoing light field—which they interpret as a spherical convolution—can be recovered from its coefficients. This task is much harder if only parts of the light field, *e.g.* a single image, is available and requires additional constraints or regularizing assumptions. For example,

any image of a sphere can be explained not only by the true illumination and BRDF, but also by a perfect mirror illuminated with that image. Experiments conducted by Fleming *et al.* [Fleming03] indicate that humans are able to match reflectance properties independent of illumination and that they do so based on prior knowledge about the behavior of natural illumination. Romeiro and Zickler [Romeiro10] therefore exploit the statistics of natural illumination to define a prior on the possible lighting L_s. Their idea is to marginalize the posterior $p(L_s, R|I) \propto p(I|L_s, R)p(R)p(L_s)$ over the lighting to obtain the probability $p(R|I)$ of the reflectance R given an image I. Computing the mean of this distribution—they in fact use an approximation—amounts to selecting a reflectance that not only explains the image for a single illumination, but for all illuminations—according to their probability. Lombardi and Nishino [Lombardi12] also consider objects of known shape but additionally recover illumination explicitly.

Laffont *et al.* [Laffont12a] reconstruct a point cloud of a scene using multi-view stereo. This serves as proxy to compute the contributions of the sun, sky, and indirect lighting based on a measured environment map. Once sun visibility and reflectance are computed for each 3D point, all this information is projected into the image and propagated over all pixels. Laffont *et al.* [Laffont12b] compute an intrinsic image decomposition but allow for varying illumination between images, which makes the technique applicable to Internet photo collections. They observe that one of the main problems with that kind of data is the unknown camera response. Lee *et al.* [Lee12] also decompose multiple images but do so for a video stream with fixed illumination. Again, their algorithm is supported by a priori known shape information—acquired from a depth camera.

3.8 Discussion

We have discussed exemplary works introducing the different ways to approach uncalibrated photometric stereo, photometric stereo with unknown reflectance, and the combination of both under the assumption of a fixed camera. Table 3.1 gives an overview over representatives of all these strategies. We observe that all of the approaches which explicitly handle non-Lambertian reflectance need either known or controlled, *i.e.* placed in a certain way, illumination. These require some sort of light calibration before the actual acquisition. The only exception is the work by Hertzmann and Seitz [Hertzmann03], which on the other hand requires the BRDF to be known in the form of an example object.

It is also apparent that most works assume a point light source, which is a severe constraint in everyday environments. Another limitation for the wide-spread employment of these techniques is that the camera response has to be calibrated in a preprocessing step. This is a great disadvantage if we consider a possible application to Internet images. We notice that only very few approaches either are invariant with respect to non-linearities or estimate the response curve as part of the reconstruction process.

In the multi-view case, we had a look at methods that exploit illumination changes and do not rely on color constancy. Again, almost none are prepared for non-linear camera responses that occur outside of the laboratory. We also observe in Table 3.2 that most approaches use a multi-step strategy that first recovers a geometric proxy

and then uses photometric information only to refine the solution in a second step. These might be iterated or followed by a fusion phase. There is no clear consensus regarding the geometry representation. This is not surprising since local and global models both have their advantages and disadvantages.

For Internet—or at least uncontrolled outdoor—images, we would like to handle arbitrary BRDFs under unknown, arbitrary illumination. Only [Treuille04] and [Chandraker13] fulfill these requirements on paper. The former needs an example object and the latter actually assumes a fixed relation of light and camera. Thus, none of the existing techniques is suitable without adaption. A further restriction in uncontrolled settings is that many approaches rely on object silhouettes. Detecting those reliably in cluttered scenes is not trivial.

Single-View Methods

Reference	Reflectance	Illumination		Strategy	Non-linear	Remarks
	model	known	type		response	
[Yuille97]	Lambertian	N	P (+ambient)	factorization + integrability	N	
[Basri01a]	Lambertian	N	A	factorization + constraints	N	
[Shi10]	Lambertian	N	P	factorization	Y	radiometric calibration (color ratios)
[Papadhimitri13]	Lambertian	N	P	perspective camera model	N	
[Shen09]	Lambertian	N	A	factorization	N	sparse points; Internet data
[Abrams12]	Lambertian	Y	P (+ambient)	optimization	Y	webcam data
[Yu13]	Lambertian	Y	A	direct least squares	N	mirror sphere needed
[Hertzmann03]	isotropic (known)	N	A	example object	Y	
[Sato94b]	diffuse+specular	Y	E	illuminant color	N	
[Goldman05]	material mixing	Y	P	non-linear optimization	N	
[Higo10]	special	Y	P	consensus/multiple constraints	Y	
[Shi12b]	isotropic	Y	P	BRDF monotonicity	Y	light distribution over hemisphere
[Zickler02]	arbitrary	C	P	Helmholtz reciprocity	N	recovers depth (if multi-view)
[Alldrin07a]	isotropic	C	P	bilateral symmetry	Y	more constraints for unique normals
[Chandraker11]	isotropic	C	P	differential images	N	circular light motion
[Aittala13]	parametric	C	E	non-linear optimization	N	almost planar targets

Table 3.1: A selection of related works for the single-view case with different assumptions about reflectance and lighting. We list the assumptions of the underlying models even though the approaches might cope with more general conditions, *e.g.* specularities, through outlier handling. Illumination can be known (Y), unknown (N), or controlled (C) and one of: point light (P), arbitrary (A), or extended (E). Some approaches can—at least in theory—handle unknown radiometric response curves (Y in column six), but most do not (N).

Multi-View Methods

Reference	Reflectance model	Illumination		Geometry representation	Optimization strategy	Remarks
		known	type			
[Zhang03]	Lambertian	N	P	DM + NM	M	
[Lim05]	Lambertian	N	P (+ambient)	DM	M	
[Vogiatzis06]	Lambertian	N	P	mesh	M	based on silhouettes
[Joshi07]	Lambertian	N	P	DM + NM	M	
[Hernandez08]	Lambertian	N	P	mesh	M	based on silhouettes
[Park13]	Lambertian	N	P	mesh + displacement map	M	based on silhouettes + multi-view stereo
[Weber02]	Lambertian	Y	P	voxel	C	
[Simakov03]	Lambertian	Y	A	DM	D	known, small object motions
[Birkbeck06]	Lambertian	Y	P	mesh	D	final fitting (Blinn-Phong)
[Yoshiyasu11]	Lambertian	Y	P	implicit surface/mesh	D	based on silhouettes
[Yang03]	diffuse + specular	N	P	voxel	C	
[Treuille04]	arbitrary (known)	N	A	voxel	C	example object
[Chandraker13]	arbitrary	N	A (fixed)	DM + NM	D	only small object motions
[Schuster10]	parametric (Cook)	Y	P	voxel	D	camera/light dome
[Yoon10]	parametric (Blinn-Phong)	Y	P	implicit surface	M	based on foreground-background separation
[Tunwattanapong13]	parametric (Ward)	Y	SH (controlled)	mesh	M	complex setup
[Zhou13]	arbitrary	Y	P (sequence)	oriented points	Q	structure from motion points to fix iso contours

Table 3.2: The most relevant related works for the multi-view case. We list the assumptions of the underlying models even though the approaches might cope with more general conditions, *e.g.* specularities, through outlier handling. Note that all techniques which operate on arbitrary BRDFs still assume isotropy. Illumination can be known or unknown (Y/N) and one of: point light (P), arbitrary (A), or spherical harmonics (SH). The geometry representation is either local (DM: depth map, NM: normal map) or global (mesh, implicit surface, voxel, or oriented points). The optimization strategy is: direct (D), multi-step (M), voxel carving/coloring (C), or propagation (Q).

Chapter 4

Calibration for Appearance Reconstruction

The typical scenario in optics is that of light entering a scene, interacting with the contained surfaces and then being observed, *e.g.*, by the human eye or a photo sensor. In our case, the measuring device is a digital camera, which transforms incoming light into discrete pixel values. So far, we have focused on appearance reconstructions that treat surface properties such as orientation or reflectance as the unknown variables. For these techniques to work, the other constituents—illumination and camera—need to be known to make the problem tractable.

Camera calibration can be divided into two parts. The first one answers the question: "Given a 3D point P, which location p on the image plane does it correspond to?". This amounts to estimating the parameters of the geometric camera model introduced in Section 2.3.1. The second aspect of camera calibration deals with the question: "Given a luminance distribution arriving at the camera, what pixel values is it transformed to and vice-versa?". This is related to the radiometric model given in Section 2.3.2. We describe some established ways of performing the first kind of calibration. Then, we present a simple model for the second one that can be applied to Internet images. We provide a first intuition about its quantitative performance and show its application in the context of visual perception which has not been considered before on this kind of data.

Estimation of light sources is an elementary building block for various computer vision and computer graphics tasks. Many techniques have been proposed that differ in the assumed lighting models, their applicability, implementation effort, and the cost of the required capturing setup. While the focus of this thesis is on general, uncontrolled input, it is often helpful in the development process to test and benchmark certain aspects, or the overall results, of an algorithm on more controlled data. For photometric techniques, that often means that the light should be a simple point light source and that its location is known very precisely. The question is how to obtain the position with sufficient accuracy in practice. We evaluate the performance of several techniques and develop a novel algorithm based on rays reflected at spherical surfaces. Compared to other approaches, we take the "relative curvature" of the spheres into account and estimate not only the direction of the light, but its 3D position.

In the end, we are mostly interested in any calibration to serve as input for a photometric reconstruction technique. It is therefore interesting to look at the interplay

of these preprocessing steps and all other potential error sources in that context. We perform a theoretical error analysis and several experiments with a diffuse sphere of known radius to answer the question: "What level of accuracy can we expect under controlled conditions?". Among others, we provide exemplary results for the response curve estimation of a consumer camera, distant point light calibration, image noise measurements, and calibrated photometric stereo.

4.1 Geometric Camera Calibration

A multitude of approaches exist to obtain the extrinsic and/or intrinsic camera parameters. They differ in the camera model, the constraints that can be exploited, and the handling of noise. Here, we summarize the steps for the kind of calibration that is most relevant in later chapters.

Intrinsic Parameters from Planar Target: Points X_i lying on a plane in 3D are related to the image positions $x_{i,j}$ by a homography H_j. If X_i and $x_{i,j}$ are known, H_j can be estimated using, for example, non-linear least squares. Zhang [Zhang99b] shows that such a homography defines two constraints on the intrinsic parameters. Thus, two images are sufficient to recover the calibration matrix K.

Extrinsic Parameters from Known Intrinsics: We assume that corresponding points $x_{i,1} \leftrightarrow x_{i,2}$ in two images are given. If the calibration matrix K is known, we can invert its effect and compute *normalized coordinates* $\tilde{x}_{i,1}, \tilde{x}_{i,2}$. The so called *essential matrix* E encodes the relative pose $[R|t]$ of the second camera with respect to the first one and is defined by

$$\tilde{x}_{i,1}^{\mathsf{T}} E \tilde{x}_{i,2} = 0. \tag{4.1}$$

Given at least eight correspondences, these constraints can be turned into a linear problem

$$Ae = 0 \tag{4.2}$$

where e contains the coefficients of the essential matrix. From the resulting E, the rotation R and translation t can be retrieved using singular value decomposition—see [Hartley06] for details. Nistér [Nistér04] even describes a way to recover the relative pose using only five corresponding points.

Intrinsic and Extrinsic Parameters from Correspondences: If the calibration matrix K is not known beforehand, it is still possible to obtain the extrinsic and intrinsic parameters. Given corresponding points $x_{i,1} \leftrightarrow x_{i,2}$, the *fundamental matrix* F defines their relation as

$$x_{i,1}^{\mathsf{T}} F x_{i,2} = 0. \tag{4.3}$$

Similar to the essential matrix, F can be recovered from eight or more of these constraints. From F, we obtain camera matrices P, but these are only unique up to a projective transformation. This can be fixed with additional constraints such as a

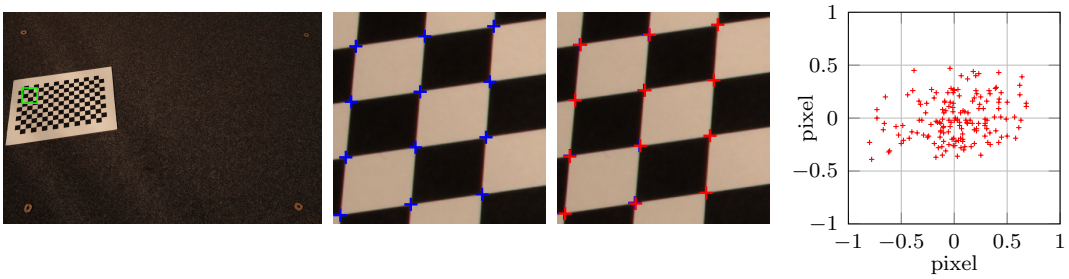


Figure 4.1: Geometric calibration. *From left to right*: An example calibration image with the checkerboard at the extreme left (closeup area indicated in green); a closeup showing the detected corners $x_{i,j}$ (blue); a closeup of the reprojected 3D points X_i (red) overlayed over the detected corners; the image space errors are below one pixel on this 5616×3744 image.

known principal point, *e.g.* $c_x = c_y = 1/2$ in Equation (2.36). Depending on the constraints, multiple images might be necessary. The details of this construction and the various types of conditions possible are beyond the scope of this thesis. We refer the reader to [Hartley06, Chapter 19] for further details. Once the final matrices $\tilde{P} = [M|Kt]$ are known, we employ a QR decomposition to find K and R as $M = KR$.

Bundle Adjustment: In practice, the steps described above are not optimal because correspondences are corrupted by noise. The results are usually just used to initialize a non-linear optimization. This *bundle adjustment* jointly optimizes the camera parameters—possibly including a distortion function d_{k_1,k_2}—and the 3D positions X_i corresponding to $x_{i,j}$:

$$\underset{P_1,\ldots,P_n,X_1,\ldots,X_m,k_1,k_2}{\arg\min} \sum_{i,j} \left\| x_{i,j} - d_{k_1,k_2}\left(P_j\, X_i\right) \right\|^2. \tag{4.4}$$

In a complete scene with multiple camera locations, we first extract the relative pose of an initial pair and then iteratively add more cameras. The 3D points are obtained by triangulating the observed correspondences $x_{i,j}$ in multiple images. Finally, this initialization is substantially improved by optimizing Equation (4.4).

In controlled settings, where intrinsic camera parameters are fixed but unknown, we first capture 30 to 50 images of a checkerboard pattern. The corners are automatically detected and then used as correspondences. Figure 4.1 shows an example after bundle adjustment and demonstrates the performance that is possible with enough input images. We only keep the highly accurate intrinsic parameters from this preprocessing step. Based on these, the calibration during the actual capturing has to consider the extrinsic parameters only and can be performed robustly even for few images.

For uncontrolled data, we apply a robust structure from motion system as described by Snavely *et al.* [Snavely06]. The registration process works by extracting feature descriptors at interest points in all images. The descriptors are matched against each other to find corresponding points in multiple images. From these, the relative pose between cameras can be computed similar to the pipeline discussed here. Again, a global bundle adjustment refines all the parameters.

4.2 Radiometric Camera Calibration

Radiometric calibration is the process of determining the non-linear relationship of scene luminance and observed pixel value given by Equation (2.44). If this relationship is fully established, we call it an *absolute calibration*. If calibration is only performed up to an unknown multiplicative factor, we call it *relative*. From relative luminances, we can only recover relative reflection properties even if the light source is known. Many inverse rendering techniques—including the approach presented in Chapter 5—accept this multiplicative ambiguity to not have to deal with absolute calibration.

In this section, we study the question: "Can we estimate absolute luminance values per pixel for everyday Internet images?". This would allow us to use community photo sites such as Flickr as "community luminance meters". We could then for example collect statistics about absolute luminance distributions in different scenes or attach true physical scales and units to reflectance reconstructions based on such images. The answer to this question is, however, also important for another reason that motivates this thesis: recreating the experience an observer has in a scene.

The human visual system reacts differently to a scene depending on the absolute luminance distribution. In low light scenarios, for example, we perceive everything in tones of gray and the acuity decreases. A camera on the other hand is geared towards producing visually pleasing images and does not take these effects into account. Pictures taken under low light conditions will lead to long exposure times but in the end produce a sharp, colorful image—disregarding any noise. This is, however, not the experience an observer would have in the actual scene, and such images therefore tend to look artificial. To really bridge the gap between computer vision, computer graphics, and perception, these aspects have to be considered.

4.2.1 Related Work

There exists a large body of literature on radiometric camera calibration, *e.g.*, [Robertson99, Robertson03, Grossberg04, Lin04, Litvinov05, Matsushita07, Kim08b]. These papers focus mostly on the calibration up to relative luminance values and typically consider a single camera. Recently, the need to make predictions of radiometric calibration on Internet data has arisen since Internet photo collections have become an input source for computer vision algorithms. Chakrabarti *et al.* [Chakrabarti09] study the imaging pipeline from scene luminance to final pixel value in this context. They propose a camera model that also encompasses the scene-dependent variation, but consider only relative luminance values. Xiong *et al.* [Xiong12] learn the relationship between scene luminance and image intensity, formulated as a Gaussian process likelihood, from large sets of training data. While the testing phase could in theory also be applied for images from the Internet, it remains to be seen how well such a training generalizes to uncontrolled test data. Kuthirummal *et al.* [Kuthirummal08] analyze the statistics of images in photo collections grouped by camera model and lens settings. They derive a prior on the joint histogram of irradiances at neighboring pixels and can then estimate response functions for other camera models in the set. Kim *et al.* [Kim12] extend previous calibration methods by modeling not only the response curve but also color space transformations and the non-linear gamut mapping applied internally by the camera. They fit parameters of this model per camera, per picture-style setting, and per white balance setting based on a large

database of controlled, but very diverse, images with raw data available. This approach can predict the relative luminance very well, but it is again unclear how well it would perform on downloaded images. Based on a 3D model of the scene, Diaz and Sturm [Diaz11a, Diaz11b] estimate the camera response for collections of photos taken by different cameras assuming a Lambertian reflectance model. Again, these just recover luminance values up to an unknown scaling factor.

Absolute camera calibration is usually considered only under controlled conditions. Martinez-Verdu *et al.* [Martinez-Verdu03] perform a series of measurements to calibrate a digital camera in the laboratory. After their colorimetric characterization, a standard camera can be used as absolute colorimeter. They discover that predicted colors from the camera differ from the true values according to an affine model. The characterization does of course not scale to the multitude of cameras that are used to take pictures on the Internet. Wueller and Gabele [Wueller07] calibrate the response functions of several digital cameras for a certain exposure level. They can then transform measurements under different exposure settings to the calibrated, absolute luminance values. Comparing these values in different scenes to those measured with a luminance meter, they find possible deviations of more than 30 % for colored objects. Brady and Legge [Brady09] also carefully calibrate a camera and use it as a luminance meter. In addition, they consider predicting the cone responses of human color perception. Their predictions for absolute luminance calibration show greater agreement with measured values than [Wueller07].

These works answer the question whether absolute luminance values can be recovered from image intensities. We would like to know if this is also true for uncontrolled Internet images. A rigorous calibration in the laboratory is not possible for that kind of data. Thus, we abandon these calibration steps and replace them with a less reliable but more general model obtained by combining several definitions from accepted camera standards.

4.2.2 Absolute Luminance from Metadata

Given a pixel value p in an image downloaded from the Internet, we would like to recover the absolute scene luminance L that was observed. We further assume that the image contains meta information about the sensor sensitivity, the focal length, and the aperture of the lens. These are part of the EXIF metadata that many images contain.

In theory, the ISO standard 12232 [ISO06] then specifies a relation of camera output to scene exposure. The proportionality factor β in Equation (2.43) equals 0.65 according to Annex B of the standard:

$$H = 0.65 \frac{L \cdot t}{N^2}. \tag{4.5}$$

We further assume the camera response curve f to be a gamma curve as defined by the sRGB standard [Stokes], which is commonly used for Internet images. The missing link between a certain absolute exposure H and an output pixel value p is then just the sensitivity γ of the sensor.

ISO standard 12232 [ISO06] includes several definitions of sensitivity. But camera vendors do not always comply with this standard exactly. Even if they do, it is usually not apparent from the available metadata which definition was used. This can lead to

greatly differing exposures for the same pixel value. A value of 200 for the ISO speed interpreted as *standard output sensitivity* implies that $H = 10/200$ gets mapped to a pixel value of 118. If it is interpreted instead as *saturation-based speed*, a pixel value 118 corresponds to an exposure of $f^{-1}(118/255) \cdot 78/200 = 0.1835 \cdot 78/200$ which is 1.43 times the previous one.

According to the EXIF standard, the sensitivity type should be recorded along with the sensitivity value. Unfortunately, many images contain only the "ISO" tag and do not specify which definition it relates to. We assume that the standard output sensitivity ("SOS") based speed S as defined in Section 7.1 of the standard can be applied for all images. Under these assumptions, the luminance L_{SOS} that produces an output value of $255 \cdot f(\hat{p}) = 118$ is given as

$$L_{SOS} = \frac{1000}{65} \cdot \frac{N^2}{S \cdot t} \tag{4.6}$$

from Equations (13, B.2) in the standard. More generally, a pixel value $p = 255 \cdot f(\tilde{p})$ corresponds to luminance

$$L = \frac{f^{-1}(p/255)}{f^{-1}(118/255)} \cdot L_{SOS} = \tilde{p} \cdot 5.5 \cdot \frac{1000}{65} \cdot \frac{N^2}{S \cdot t} \,. \tag{4.7}$$

Since we typically deal with color images, we cannot use this formula directly. Instead, we compute grayscale values by combining the RGB channels according to Section 6.3.3 in ISO 12232:

$$\tilde{p} = 0.2125 \cdot f^{-1}\left(\frac{R}{255}\right) + 0.7154 \cdot f^{-1}\left(\frac{G}{255}\right) + 0.0721 \cdot f^{-1}\left(\frac{B}{255}\right). \tag{4.8}$$

We tested the proposed model in a controlled environment by capturing a color checker with a Canon EOS 5D camera and simultaneously acquiring measurements with a luminance meter. It turned out that the predicted luminances according to Equation (4.7) deviate approximately ±30% from ground truth. This matches with the findings by Wueller and Gabele [Wueller07], who even calibrate the camera's response.

4.2.3 Performance on Internet Images

The model we just described relies on images to be standard conforming. It is unclear how well it performs on Internet images since they might be edited, contain false information, or because camera manufacturers do not follow the standard. Still, a lot of images should roughly comply with this model. The interesting question is therefore not so much whether we can estimate per-pixel luminance values but rather *how well* we can do it.

To answer this question, we need a large amount of images with ground truth luminance values. In a controlled environment, Wueller and Gabele [Wueller07] use a luminance meter to measure the scene and then compare to a calibrated camera. This does not match well with the variations in Internet images, which are taken by many different camera models, show different scenes, or are possibly post-processed. Instead, our approach is to use images that were actually downloaded from platforms such as Flickr and Google images. Measuring the ground-truth with a luminance meter for

these images might work for a frequently photographed indoor scene without many luminance changes (*e.g.* an exhibit in a museum). But preferably, we would like to even dispense with the need of measuring the ground truth and instead use a calibration target with known luminance.

Motivated by Ansel Adams [Adams83], we propose to use the Moon as such a target. Our hope is that insights gained on lunar images from the Internet transfer to the more general classes. In the following, we first explain how the luminance of the Moon can be computed from its phase angle and how we recover this angle automatically from images. We then compare the computed luminance with predictions obtained through metadata as described in Section 4.2.2.

How Bright is the Moon?

The apparent brightness of the Moon has been of interest for centuries since it is the second brightest object in the sky after the Sun. We will only mention some selected works. Ellis [Ellis66] combines previous results about the relative intensity of the Moon with atmospheric extinction to derive tabulated intensities for varying zenith angle and lunar phase. These values are relative to the intensity of the full moon and cannot be used directly for absolute calibration.

In 1994, the Clementine mission produced high resolution images of the lunar surface from a polar orbit. This enabled an updated model of the lunar albedo at various wavelengths and a study of the so called *opposition effect* by Buratti *et al.* [Buratti96]. The opposition effect leads to an apparent increase of albedo for small phase angles, which is also studied by Hapke [Hapke66]. Using Hapke's work, Jensen *et al.* [Jensen01] assemble a model of the complete night sky including the Moon, which can be used for realistic image synthesis. Their formulation is centered on irradiance, which we transform into radiance/luminance at a single camera pixel. This is the inverse of Kieffer and Stone's [Kieffer05] approach to processing lunar radiance measurements in the context of the ROLO (RObotic Lunar Observatory) program [USGS]. The ROLO program by USGS is aimed at providing radiometric calibration for space-based imaging instruments, using the Moon as a reference source. We follow this idea but apply it in a totally different context: We assume earth-based consumer cameras instead of specialized space equipment and do not know the actual time of capture or the observer's location.

Finally, in working with Internet images, we also encountered pictures of lunar eclipses since these draw special attention. Hernitschek *et al.* [Hernitschek08] observe eclipses with a light meter and devise a model for the luminance of the Moon while in the shadow of the earth. Since this is a special case, we do not use images of lunar eclipses and consider them as outliers in our derivations.

A Model of Lunar Luminance

We make use of the lunar illumination model in [Jensen01] that relates solar irradiance on the Moon $\tilde{E}_{sm}$ to the irradiance on Earth arriving from the Moon $\tilde{E}_m$ depending on the phase angle φ:

$$\tilde{E}_m(\varphi) = \frac{2}{3} \cdot \frac{C}{\pi} \cdot \omega_M \cdot \tilde{E}_{sm} \cdot (1 - \sin(\varphi/2) \tan(\varphi/2) \log(\cot(\varphi/4))) \,. \tag{4.9}$$

This formula ignores the earthshine component which is only relevant around a new moon. The average albedo C is set to 0.072 and ω_M is the solid angle of the complete lunar disk as observed from the earth. From our camera model we obtain luminance values instead of irradiances. Exchanging radiometric quantities with photometric ones in Equation (4.9) gives us the illuminance E_m. The illuminance integrates incoming luminance L over the whole hemisphere:

$$E = \int L(D)\langle v,\ D\rangle d\omega_D \tag{4.10}$$

where v is the orientation of the sensor. Assuming that L is constant for all directions D which correspond to the illuminated portion of the Moon Ω_φ and zero everywhere else, we obtain

$$E_m(\varphi) = L \int_{\Omega_P} \langle v,\ D\rangle d\omega_D. \tag{4.11}$$

We further assume that the sensor is oriented towards the Moon, which lets us approximate $\langle v,\ D\rangle \approx 1$ and leads to

$$E_m(\varphi) \approx L \int_{\Omega_\varphi} d\omega_D = L \cdot \omega_\varphi. \tag{4.12}$$

The solid angle ω_φ can be computed from the average radius $r_M = 1737\,\mathrm{km}$ of the Moon and its average distance from Earth $d = 384\,400\,\mathrm{km}$:

$$\omega_\varphi = P(\varphi) \cdot \frac{\pi r_M^2}{d} = P(\varphi) \cdot \omega_M \tag{4.13}$$

where $P(\varphi)$ is the percentage of the lunar disc that is illuminated as seen from the earth. P depends on the phase angle as follows (*cf.* [Lun]):

$$P(\varphi) = 0.5 \cdot (1 + \cos(\varphi))\,. \tag{4.14}$$

Thus, our final model of lunar luminance is

$$L_m = \frac{E_m}{\omega_P} = \frac{2}{3} \cdot \frac{C}{\pi} \cdot \frac{2\ E_{sm}}{1 + \cos(\varphi)} \cdot (1 - \sin(\varphi/2)\tan(\varphi/2)\log(\cot(\varphi/4)))\,. \tag{4.15}$$

We use a value of $1.338 \times 10^5\,\mathrm{lm/m^2}$ for the illuminance of the Sun E_{sm} at a distance of one astronomical unit ($1\,\mathrm{au} = 149\,597\,870.7\,\mathrm{km}$) as in the CIE sky model [Darula02]. We ignore the minor deviations of the actual position of the Moon from 1 au.

Evaluation

We downloaded an initial set of about 13 000 images from Flickr with tags “moon”, “full”, “night”, and “sky”. Those without the necessary EXIF data were automatically removed. Many images show artistic works or landscape shots with the Moon as a small, over-exposed disk. We removed most of these by comparison with an actual photograph of the Moon using the PictureRelate [Walthelm] tool. The remaining 800 images were uploaded to Amazon Mechanical Turk where humans annotated them with a bounding box of the Moon.

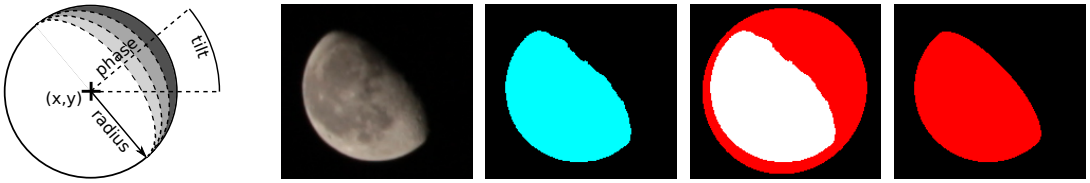


Figure 4.2: *Left to right*: The geometry of lunar illumination coverage, an example image of the Moon, the binary mask after thresholding, initial coverage (*red*) assumed as initialization for the optimization (image mask overlayed in white), estimated coverage after optimization ($\varphi = 70°$).

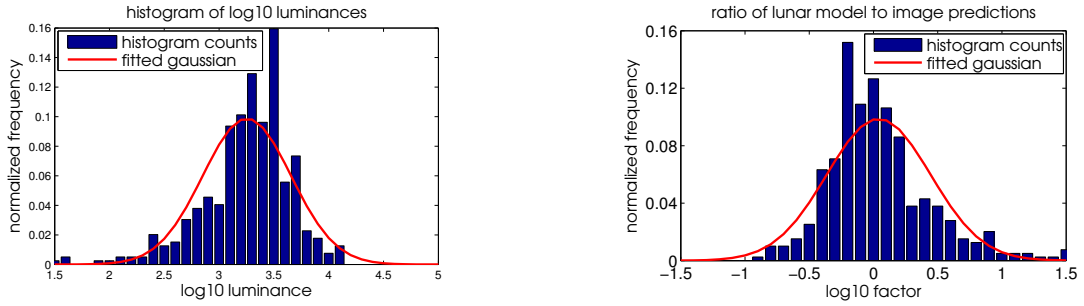


Figure 4.3: *Left*: Histogram (*blue*) of predicted $\log_{10}$ luminances from image metadata ($\mu = 3.25, \sigma = 0.4$) with corresponding Gaussian overlayed in *red*. *Right*: Histogram of differences between predictions from both models ($\mu = 0.04, \sigma = 0.4$).

Each image was then transformed according to Equation (4.8) and thresholded at $p = 77$ to create a binary mask of the Moon. We need the phase angle to evaluate the model in Equation (4.15) and compute luminances in each image. This angle can be estimated from a geometric model of the illumination coverage of the lunar disk. The model is illustrated in Figure 4.2 and accounts for unknown pixel coordinates of the center (x, y), the radius, a tilt angle, and the phase. Its output is the set of pixels that correspond to the illuminated part of the Moon (*red* in Figure 4.2). We run an optimization that fits this model to the binary mask in each image and returns the estimated lunar phase angles. For the reference image shown in Figure 4.2, we know the exact date and position of capture. Comparing the estimated phase angle of $\varphi = 70°$ with data provided by NASA's HORIZONS system [NASA] $\varphi = 65°$ yields a deviation of about $27\,\mathrm{cd/m^2}$ which is negligible for our purposes. Once φ is known, we evaluate Equation (4.15) for each image to compute the luminance values L_{lunar}.

Applying Equation (4.7), we can also compute luminance estimates based on the actual pixel values and metadata. We first exclude pixels with $p \geq 250$ as overexposed and then predict per-pixel luminance values within the mask. These are averaged to yield the predicted luminance of the Moon based on image measurements L_{image}. If more than 1 % of the pixels are overexposed, we remove the image. Figure 4.3 (left) displays a histogram of the image-based predictions greater than $30\,\mathrm{cd/m^2}$ for the remaining 395 images.

Finally, we compare the predicted luminances L_{image} with those computed from the lunar model by taking the ratio L_{lunar}/L_{image} for each image. The respective histogram of $\log_{10}$ ratios is plotted in Figure 4.3 (right). We observe that the mean ratio is close to zero, which implies that both predictions agree quite well most of

the time. The ratios $10^{-0.4} \approx 0.4$ and $10^{0.4} \approx 2.5$ corresponding to plus and minus one standard deviation are, however, rather large. Note that these numbers include all variations and unmodeled effects in the complete pipeline. Common examples are images with cloudy or not completely black sky, e.g., during dusk.

Limitations and Shortcomings

Like any comparison that is based on a theoretical model instead of concrete measurements, these results are subject to the accuracy of the underlying model. If the model did not fit reality, then a close match of the predictions to the model would not necessarily indicate a good prediction of the real world. We have to trust in the model of Jensen *et al.* [Jensen01] and our adaptions to be a reasonable approximation of reality. An obvious shortcoming is the neglect of any atmospheric effects that attenuate the incoming luminance.

At sea level, the extinction due to Rayleigh scattering R_{Ray} and aerosol scattering R_{aer} can be approximated according to Green [Green92] in its simplest form as

$$F = 2.512^{R_{Ray}+R_{aer}} \cdot A = 2.512^{0.1451+0.12} \cdot A = 1.2766\ A\ . \tag{4.16}$$

The air mass A is defined as 1 when looking straight up (zenith) and can be calculated from the zenith angle ϕ of the Moon at the observer position:

$$A = (\cos(\phi) + 0.025\ \exp(-11\ \cos(\phi)))^{-1}\ . \tag{4.17}$$

The final observed luminance will be $L = L_m/F$. Thus, neglecting atmospheric effects at sea level introduces an erroneous factor of 1.28 in the best case ($\phi = 0°$) and of 7.2 in the worst case ($\phi = 80°$, larger angles excluded because of occlusion at the horizon). These errors decrease, however, with increased altitude of the observer. Since we do not know where and when a downloaded image was taken, we cannot deduce the zenith angle or the approximate extinction coefficient F.

We therefore have to be careful with conclusions drawn from the comparison above. They should be seen as a first attempt at quantifying the errors of the radiometric model presented in Section 4.2.2. To fully answer the question of how well absolute luminance can be predicted on Internet images using metadata, a better model of the Moon is needed. In the future, we expect more Internet images to be available with geo-information and time stamps. This would allow us to extend the lunar model with atmospheric attenuation. Nevertheless, an accurate computation of lunar luminance stays a challenging problem. Perhaps a better way to assess the prediction of absolute luminance is to use Internet images of a scene with constant luminance, *e.g.* in a museum, or taken at the same time, *e.g.* the inauguration of the president of the United States on the steps of the Capitol.

4.2.4 Application in Perception

We have seen how to reconstruct absolute luminance values from a model based on metadata and discussed issues concerning a quantitative evaluation in the previous sections. How accurate we actually have to be depends, however, on the use case and context. As an application of absolute luminance in images, we now study perceptual effects in human vision.

Cameras aim at reproducing images an observer has seen in photopic conditions. However, important changes in perception occur in the mesopic and scotopic range of low light illumination. They depend on the absolute adaption state of the eye's receptor cells. A camera will therefore fail to produce authentic images under such conditions. We show how absolute luminances can be employed to re-render night images in a perceptually more plausible way.

Perceptual Effects

Several effects can arise in low light conditions and have been investigated in the context of rendered high dynamic range scenes. Spencer *et al.* [Spencer95] study glare arising from scattering in ocular media and diffraction at the iris. Ward *et al.* [Ward97] as well as Ferwerda *et al.* [Ferwerda96] also model glare, but add color sensitivity and visual acuity. Their focus is on employing local contrast thresholds for dynamic range compression, which depend on absolute luminance. Durand and Dorsey [Durand00] extend on this work and use a global adaption level per image. These examples demonstrate the importance of absolute luminance values for tone mapping and the simulation of perceptual effects. We will focus on two such effects which depend on the absolute adaption level of the eye.

Visual Acuity: Acuity expresses the spatial resolution of the visual system. This can also be interpreted as sharpness and is typically measured in cycles per visual degree. For human vision, a basic test from ophthalmology is, for example, to have the patient read letters of decreasing size at a fixed distance. It has been shown by Shlaer [Shlaer37] that performance in these tests depends on the adapting luminance. Under moonlight conditions, acuity drops significantly and spatial detail is lost even for healthy observers.

Ward *et al.* [Ward97] derive the following relation of luminance to maximal resolvable frequency from Shlaer's data:

$$\nu(L) = 25.72 + 17.25 \cdot \operatorname{atan}\left(1.4\ \log_{10}(L) + 0.35\right). \tag{4.18}$$

To simulate this effect, we blur the image with a low-pass filter that removes frequencies above $\nu(L)$. Note that the adaption luminance is not constant over the whole image, but may change locally. Different levels of blur will be appropriate in different regions of the image. We borrow from Krawczyk *et al.* [Krawczyk05] and use the same kernels as in their high dynamic range work.

Rod and Cone Activity: Another effect that becomes apparent in low light is the loss of color. In the scotopic range of about $10^{-6}\,\mathrm{cd/m^2}$ to $10^{-2}\,\mathrm{cd/m^2}$, the cone cells in the human eye are inactive and vision is solely based on the rod photoreceptor cells. Again, cameras mimic the human eye under photopic conditions when the three types of cones would provide a perception of color.

The rods have a spectral sensitivity that differs from the luminosity function of the cones, which is usually applied in photometry as discussed in Section 2.1.2. Without knowing the spectral distribution of incoming light, it is not possible to directly translate one into the other. We therefore neglect this effect, use photopic luminance values

in both cases, and implement the loss of color perception by blending a grayscale image and a colored one. The blending coefficient simulates the decrease in rod activity with increasing adaption luminance. Again, these coefficients can vary spatially over the image.

Ferwerda *et al.* [Ferwerda96] propose a linear blending weight for the mesopic range, whereas Durand and Dorsey [Durand00] employ the ratio of two linear functions of adaption luminance. We use the same method as Krawczyk *et al.* [Krawczyk05], who define the blending weight as

$$\sigma(L) = \frac{0.04}{0.04 + L} \,, \tag{4.19}$$

effectively disabling rod contribution for luminances above $10\,\mathrm{cd/m^2}$.

Implementation

All works mentioned above, apart from [Krawczyk05], consider these effects for tone mapping of computer generated high dynamic range renderings. With the availability of high dynamic range photographs through multiple exposure algorithms, tone mapping has also been applied to real scenes. For example, Reinhard *et al.* [Reinhard02] use relative luminance values and a *key value* that adjusts the overall brightness. They do not explicitly model any perceptual effects. These are then introduced to photographic tone mapping of *high dynamic range videos* by Krawczyk *et al.* [Krawczyk05]. We show that effects such as reduced acuity in low light conditions can also be employed with the more common *low dynamic range images* to increase their "naturalness".

We assume an sRGB image as input and first transform it into the xyY space. Using Equation (4.7), we scale the Y channel to obtain absolute luminance values. We then build a Gaussian pyramid of successively blurred luminance images similar to Krawczyk *et al.* [Krawczyk05] but without their approximation of kernels since we are not concerned about real-time performance. Based on the absolute luminance and Equation (4.18), we select the appropriate pyramid level for each pixel and obtain a blurred luminance map L_{ac}.

We then split this map into a scotopic part $L_s = \sigma(L_{ac}) \cdot L_{ac}$ and a photopic one $L_p = (1 - \sigma(L_{ac})) \cdot L_{ac}$ according to the rod activity σ. It is important to note that the blending coefficient depends on absolute luminances. In the next step, we bring the results back into the range of initial Y values by reversing Equation (4.7) and obtain the scotopic Y_s and photopic Y_p. Y_p is transformed into the XYZ space using the original chromaticities xy and then added to the respective grayscale image obtained from Y_s and $x = 1/3, y = 1/3$. Finally, we convert the image to the output space which is again sRGB.

Figure 4.4 shows the different images occurring in this algorithm. The original was taken in a dark capture lab with the white glove still recognizable by a human. The color checker became visible after a short period of dark adaption. The sharp and colored camera picture with an exposure time of 15 s is very far from the experience a human would have in this setting.

Figure 4.4: Stages of perceptual simulation. *Top*: The original image taken with a long exposure time shows bright colors and sharp details (left). We remove spatial frequencies that could not be resolved by an observer under low light conditions (right). *Bottom*: Rod responses are simulated through a grayscale image (left), which is then blended with the colored one according to local rod activity. The final result (right) shows both effects combined.

Results on Internet Images

To test our simulation of human perception, we downloaded images from Flickr that were taken in low light, but still look sharp and colorful due to long exposure times. It is immediately apparent that the original images in the left column of Figure 4.5 look artificial. This is of course intended and an artistic choice by the photographers, but we are interested in how a human would actually have perceived the scene. In our results in the right column, the loss of acuity can be well observed on the surface of the rock in the first row, at the air vent in the second row, and at the fence in the third row (see also the closeups in Figure 4.6). Similarly, colors are much less saturated. Note especially the green plants and tufts of grass in the third row and the overall appearance in the first row. In general, our images look more plausible with regard to a night time scene than the original long-exposure shots. Of course, daylight images should not be affected by our simulation, and we show one such example in Figure 4.7.

In these examples, we are dealing with a range of luminances from starlight at about $0.001\,\mathrm{cd/m^2}$ to clouded sky illumination of $1000\,\mathrm{cd/m^2}$. Figure 4.8 shows false color images of the absolute log luminances in all four images. To study the impact that possible false estimates of the luminance factor could have on overall image

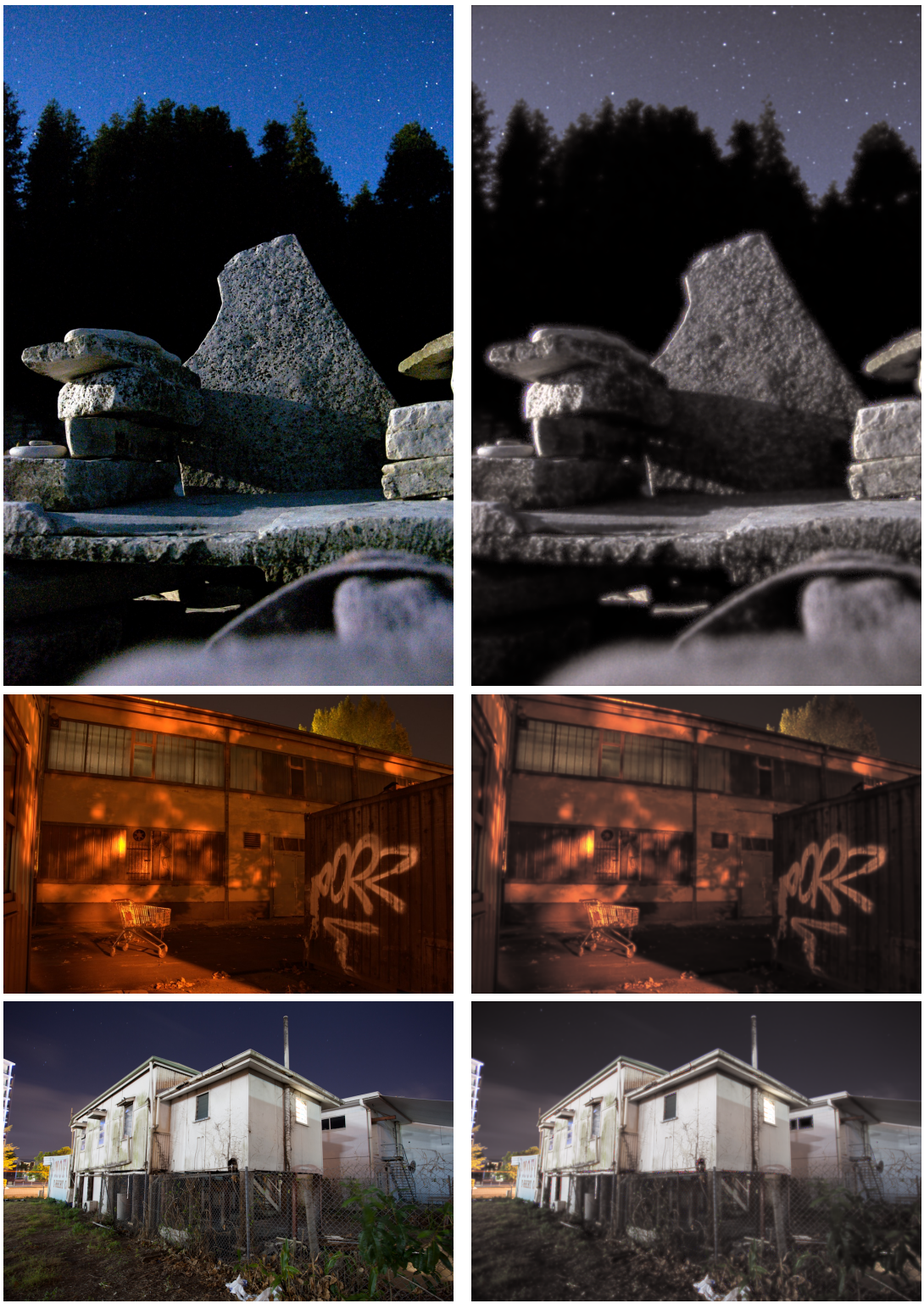


Figure 4.5: Results. *Left*: The original, unprocessed images (rows 1-3: Flickr users Joselito Tagarao, Markus Lehr, Wayne Grivell). *Right*: Results show-casing reduced acuity and color perception in low light conditions.

Figure 4.6: Closeups. *Left*: The air vent in the original image by Markus Lehr and the perceptual re-rendering. *Right*: The fence in the original image by Wayne Grivell next to our result.

Figure 4.7: Results on a daylight image. *Left*: The original image. *Right*: Results after applying the pipeline. As expected, the daylight picture is not affected.

appearance, we artificially introduced an additional scale factor α which we varied among $0.2, 0.5, 1.0, 2.0, 5.0$. Figure 4.9 shows that deviations by a factor of 5 are clearly discernible. For $\alpha = 2$, we can make out a difference, but the overall impression of a low light scene is still preserved. Thus, in this use case, deviations of one f-stop might still be acceptable.

Figure 4.8: Colormapped $\log_{10}$ of absolute luminance. Images from left to right correspond to rows in Figure 4.5 and Figure 4.7. The respective logarithmic mean of estimated luminances is: $0.0003\,\mathrm{cd/m^2}$, $0.014\,\mathrm{cd/m^2}$, $0.018\,\mathrm{cd/m^2}$, $92.86\,\mathrm{cd/m^2}$. The estimated luminance factor is: $0.087\,\mathrm{cd/m^2}$, $0.576\,\mathrm{cd/m^2}$, $0.407\,\mathrm{cd/m^2}$, $1128.23\,\mathrm{cd/m^2}$.

Figure 4.9: Impact of possible false estimates. *Left to right*: results for an additional scaling factor of $0.2, 0.5, 1, 2, 5$ in the luminance estimation.

4.3 Geometric Point Light Source Calibration

Known (point) light source positions are the basis of many shape or reflectance recovery techniques such as photometric stereo. Most of these algorithms are still designed to operate in a controlled indoor environment, *e.g.* a scientific laboratory or a movie set. If space allows to place the light sufficiently far away, many light estimation techniques [Dosselmann13, Wang02, Wong08, Zhou02] assuming infinitely distant illumination are available to recover its overall direction. If the light cannot be placed sufficiently far away, however, its direction is not constant for every scene point, and the irradiance falls off with the square of the distance.

In this section, we propose two new methods to recover the position, and thus also the direction, of a point light source. For the first setup, we require a single image and two or more mirror spheres that are placed in the scene. The reconstruction is then based on minimizing the image-space error of the light highlights reflected from the spheres. For the second technique, we require several images of the scene that directly observe the light source and a couple of feature points—the sphere centers in our case. We then reinterpret the problem of light source estimation as a 3D reconstruction task, which we solve with techniques from Section 4.1.

To analyze their strengths and weaknesses, we study the performance of both techniques with respect to one another and to the traditional approach of ray intersections. Selecting one method in practice will not depend on accuracy alone but also on the effort during capture and its implementation cost. Finally, and in contrast to other works in this area, we analyze the impact of the spatial arrangement and the number of spheres on robustness and accuracy of the solution.

4.3.1 Related Work

There is a large body of literature on light source estimation. The approaches differ in accuracy, capture setup, lighting model, additional constraints, and in the intended application, *e.g.*, renderings in augmented reality, shape reconstruction, or image-based relighting. Some works exploit cast shadows [Panagopoulos11], sample the complete incoming light-field [Sato99, Kanbara04], or estimate the light source from stationary images [Winnemoeller05]. Approaches that minimize an intensity error compare actual images of a scene with known geometry and reflectance against renderings with the current light estimate [Hara05, Weber01, Xu08]. These are of course unsuited in our context where the shape is to be reconstructed in a later step. We avoid this problem by placing target objects with simple, known geometry in the scene.

We narrow down the discussion of related techniques to a selection in the subfield of point light source estimation based on spherical target objects. Some of the ideas used for recovering infinitely distant illumination can be readily used to estimate positions of near point light sources if applied to multiple spheres. For example, Masselus *et al.* [Masselus02] demonstrate that once light directions are known with respect to several scene points, the corresponding rays can be intersected to yield a light position. In particular, they use four diffuse spheres to obtain the directions. They do not perform a quantitative evaluation on real images. Powell *et al.* [Powell01] show that obtaining the respective light directions is especially easy for reflective spheres at known positions. They use two spheres in a special setup with a fixed baseline of 11 cm and assume that reflection points in 3D correspond accurately to detected image highlights. The framework of Zhou *et al.* [Zhou04] is based on images of specular spheres placed at different locations to triangulate an area light source. They do not evaluate the impact the number of images has on their results. Nayar [Nayar89a] uses mirroring spheres for 3D reconstruction and shows a strong relation to multi-view stereo. He evaluates his reconstruction framework in the context of light source triangulation. All these methods assume that highlights on the spheres are detected accurately in the image.

Aoto *et al.* [Aoto12] are the only ones to consider the reprojection error for triangulation of near light sources. Their setup consists of a hollow glass sphere with known position and radius. Due to the inherent difficulty of computing the 3D position of the reflection point on the surface of the sphere, the authors exploit a characteristic of epipolar geometry to triangulate the light source using the two highlights on the front and the back side of the sphere. This limits their approach to a small baseline defined by the diameter of the sphere and consequently yields unstable results for distant light sources. In contrast, our approach enables us to use an arbitrary baseline, which cannot be achieved with a single glass sphere.

4.3.2 Approaches

Light reflected at a mirroring surface creates a distinct highlight that can be easily detected in the observed image but which is not error-free. We will first present the most common way of obtaining the light position L from such data. It works by casting rays towards the highlight on at least two spheres and intersecting the reflected rays in 3D. This is what we call the *forward calibration* as rays are shot *forward* from the camera. The triangulation problem is well known in the context of image-based scene

reconstruction, *cf*. [Hartley97], and it is preferable to minimize the reprojection error instead of computing the closest point to all rays in 3D space.

We will then introduce our new *backward calibration* which evaluates the error in image space by tracing rays from the light source to the spheres and back to the camera. Afterward, a third method which directly triangulates the light position with high accuracy is presented. For all explanations, we assume that the sphere position S and radius r are known in the camera coordinate system. We will discuss ways to obtain the sphere position from the image and r alone in Section 4.3.4.

Forward Calibration

The commonly employed method to perform light calibration is by finding the closest point in 3D to a series of rays. Masselus *et al*. [Masselus02] obtain these rays for diffuse spheres by inverting a linear shading model similar to Equation (2.28). For mirror spheres, the typical approach is to shoot rays u through the observed highlight pixels [Powell01, Nayar89a]. It is then straightforward to solve the quadratic polynomial

$$r^2 = \|\lambda u - S\|^2 \tag{4.20}$$

to obtain the intersection $R = \lambda u$ with a known sphere. Reflecting at the intersection normal $N = (R - S)/r$ gives the ray $v = u - 2(N^\mathsf{T} u)N$ originating at R.

Once the rays v toward the light source are known, the light position L is given as the position that minimizes the squared distance to all rays. Projecting $\hat{L} := L - R$ orthogonally onto the ray v yields a decomposition $\hat{L} = \hat{L}_\parallel + \hat{L}_\perp$ with $\hat{L}_\parallel = (v^\mathsf{T}\hat{L}) \cdot v$. The orthogonal distance $\|\hat{L}_\perp\|$ can then be expressed with matrices $A = (\mathrm{id} - vv^\mathsf{T})$ and $b = AR$ as

$$d = \|\hat{L} - \hat{L}_\parallel\| = \|(\mathrm{id} - vv^\mathsf{T})\hat{L}\| = \|A \cdot L - b\|. \tag{4.21}$$

We minimize the squared distance to all rays simultaneously:

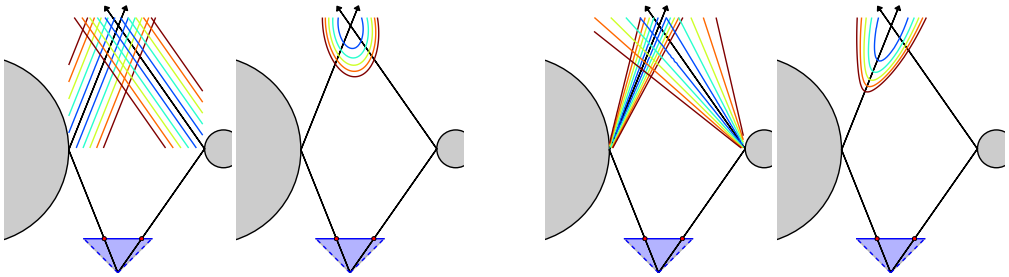
$$\arg\min_L \sum_{i=1}^{n} d_i^2 = \arg\min_L \|(A_1^\mathsf{T}, \ldots, A_n^\mathsf{T})^\mathsf{T} \cdot L - (b_1^\mathsf{T}, \ldots, b_n^\mathsf{T})^\mathsf{T}\|^2. \tag{4.22}$$

Figure 4.10 (left) shows a visualization of this energy and its individual contributions for two exemplary spheres.

In practice, the pixel positions of the detected highlights will deviate slightly from the ground truth. We can interpret the error distribution—or its logarithm—as a probability for the true location of the light given the observations. The maximum is attained at the intersection point, and both spheres contribute equally. A small error in the highlight detection of the smaller sphere, however, affects the overall result much more than deviations on the larger sphere. We would like this to be represented in the probability distribution. The backward calibration explained in the next section automatically reduces the contribution of the smaller sphere in such cases.

Backward Calibration

In the case of the backward calibration, our idea is to optimize the light source position by minimizing the projection errors of the reflections R. Let $R_i(L)$ be the reflection



(a) forward error (b) backward error

Figure 4.10: 2D example. A visualization of the error distribution for the forward (a) and backward (b) case. The left illustration in both cases shows the individual contributions of each sphere whereas the right contains the summed error. The red dots indicate detected highlights on the image plane, and colored lines are the iso-contours of the energy.

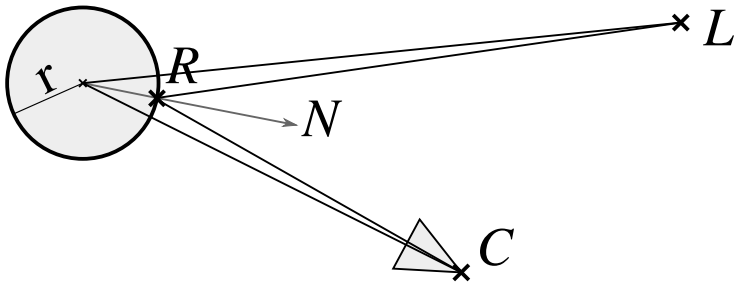


Figure 4.11: Reflection geometry. The difficulty in backward calibration lies in determining the point of reflection R that is generated from light source L and reflected into camera C.

of the light L in sphere i, and H_i the detected highlight. The task is to minimize

$$\arg\min_{L} \sum_{i} \|\pi(KR_i(L)) - H_i\|^2 \tag{4.23}$$

where K is the calibration matrix of the camera and π a projection operator. This case is more difficult because we do not know which ray to intersect with the sphere. To our knowledge, it has not been studied for light calibration with a general constellation of mirror spheres. Again, we assume the sphere position S and radius r to be known in the camera coordinate system. The challenge is to compute the highlight position R in 3D for a given light source position L. Figure 4.11 illustrates this situation.

We first translate the camera coordinate system into the known sphere center S which yields a camera position $\widetilde{C} = -S$. For a light source position $\widetilde{L} = L - S$, a unique point $\widetilde{R}$ on the surface of the sphere reflects towards the camera—assuming that a reflection exists at all. Note that $\widetilde{R}$ does not in general bisect the angle between $\widetilde{L}$ and $\widetilde{C}$ but rather the angle between $\widetilde{L} - \widetilde{R}$ and $\widetilde{C} - \widetilde{R}$. Aoto *et al.* [Aoto12] do not compute the reflection point $\widetilde{R}$ for a general arrangement and only remark on the difficulty of that problem. We show how this can be solved and review the geometric reasoning by Eberly [Eberly], which leads to a quartic equation.

If $\widetilde{L}$ and $\widetilde{C}$ are not parallel, we can use them as basis vectors and decompose the

unknown point as $\widetilde{R} = x\widetilde{C} + y\widetilde{L}$. A first constraint is then given by the radius r as

$$r^2 = \widetilde{R}^\mathsf{T}\widetilde{R} = x^2\widetilde{C}^\mathsf{T}\widetilde{C} + 2xy\widetilde{C}^\mathsf{T}\widetilde{L} + y^2\widetilde{L}^\mathsf{T}\widetilde{L}. \tag{4.24}$$

We obtain a second constraint by reflecting $\widetilde{C}$ across the line described by $\widetilde{R}$:

$$\widetilde{C}' = 2\frac{\widetilde{C}^\mathsf{T}\widetilde{R}}{\widetilde{R}^\mathsf{T}\widetilde{R}}\widetilde{R} - \widetilde{C} = 2\frac{x\widetilde{C}^\mathsf{T}\widetilde{C} + y\widetilde{C}^\mathsf{T}\widetilde{L}}{r^2}\widetilde{R} - \widetilde{C} =: 2\alpha\widetilde{R} - \widetilde{C}. \tag{4.25}$$

The reflected point $\widetilde{C}'$ lies on the line from $\widetilde{R}$ to $\widetilde{L}$. Thus, $\widetilde{C}' - \widetilde{R}$ is parallel to $\widetilde{L} - \widetilde{R}$:

$$0 = (\widetilde{L} - \widetilde{R}) \times (\widetilde{C}' - \widetilde{R}) = (\widetilde{L} - \widetilde{R}) \times \left((2\alpha - 1)\widetilde{R} - \widetilde{C}\right) \tag{4.26}$$

$$= (2\alpha - 1)x\widetilde{L} \times \widetilde{C} - \widetilde{L} \times \widetilde{C} + y\widetilde{L} \times \widetilde{C} \tag{4.27}$$

$$= (2\alpha x - x - 1 + y)\widetilde{L} \times \widetilde{C}. \tag{4.28}$$

Since $\widetilde{L}$ and $\widetilde{C}$ were assumed not to be parallel ($\widetilde{L} \times \widetilde{C} \neq 0$), it follows that

$$0 = 2\alpha x - x - 1 + y \tag{4.29}$$

$$= 2r^{-2}(x\widetilde{C}^\mathsf{T}\widetilde{C} + y\widetilde{C}^\mathsf{T}\widetilde{L})x - x - 1 + y \tag{4.30}$$

$$= 2r^{-2}\widetilde{C}^\mathsf{T}\widetilde{C}x^2 + 2r^{-2}\widetilde{C}^\mathsf{T}\widetilde{L}xy - x + y - 1. \tag{4.31}$$

Equation (4.24) and Equation (4.31) are two polynomials in the coordinates of $\widetilde{R}$. Introducing $c := r^{-2}\widetilde{C}^\mathsf{T}\widetilde{C}$, $b := r^{-2}\widetilde{C}^\mathsf{T}\widetilde{L}$, $a := r^{-2}\widetilde{L}^\mathsf{T}\widetilde{L}$, and separating y in Equation (4.31) yields

$$y = \frac{1 - 2cx^2 + x}{2bx + 1}. \tag{4.32}$$

We insert this result into Equation (4.24) and reorder:

$$\begin{aligned} 0 = {} & 4c(ac - b^2)x^4 - 4(ac - b^2)x^3 \\ & + (a + 2b - 4ac + c)x^2 + 2(a - b)x \\ & + a - 1. \end{aligned} \tag{4.33}$$

We know that this fourth order polynomial equation has at least one real solution because the reflection exists in all non-degenerate cases. We obtain it with a standard technique (see [Bronstein08]) which instead computes the roots of

$$x^2 + \frac{\beta + A}{2}x + \left(z + \frac{\beta z - \delta}{A}\right) \tag{4.34}$$

with $\beta = -1/c$, $\delta = \frac{a-b}{2c(ac-b^2)}$, $\gamma = \frac{a+2b-4ac+c}{4c(ac-b^2)}$, $A = \pm\sqrt{8z + \beta^2 - 4\gamma}$, $e = \frac{a-1}{4c(ac-b^2)}$, and z any real solution of the cubic equation

$$8z^3 - 4\gamma z^2 + (2\beta\delta - 8e)\,z + e(4\gamma - \beta^2) - \delta^2 = 0. \tag{4.35}$$

We pick the positive solution x of Equation (4.34) which corresponds to $\widetilde{R}$ lying *between* $\widetilde{L}$ and $\widetilde{C}$. With y from Equation (4.32), the reflection point is given as $\widetilde{R} = x\widetilde{C} + y\widetilde{L}$.

Finally, we translate back into the camera coordinate system and obtain $R = \widetilde{R}+S$. The projection of this point into the image contributes to the overall error according to Equation (4.23) and is visualized in Figure 4.10 (right). We then solve the resulting non-linear least squares problem using the Ceres [Agarwal] optimization library. In our tests, we did not observe the optimization getting stuck in local minima when restarting with different initial conditions.

Direct Light Position Triangulation

Another way of obtaining the light source position is to include the light directly in the images of the scene. This is often not applicable if the light source is far away from the scene. If feasible, however, this method yields impressive results as we will show in our evaluation in Section 4.3.5. A related approach has been proposed by Frahm *et al.* [Frahm05] in the context of augmented reality with light source estimation. In contrast to their approach, we do not use light tracking but robust camera calibration with bundle adjustment.

First, we exploit the 2D coordinates $p_{i,j}$ of the spheres to estimate the extrinsic camera parameters as described in Section 4.1. We then perform an initial triangulation of the sphere centers $P_i \in \mathbb{R}^3$ and light source position through direct ray intersection similar to Equation (4.21). Finally, all these positions are substantially improved using standard bundle adjustment, which reduces the global error over all images:

$$\sum_j \Big(\sum_i \|\pi_j(K_j P_i) - p_{i,j}\|^2 + \|\pi_j(K_j L) - h_j\|^2 \Big) \tag{4.36}$$

where h_j is the pixel position of the light source.

In order to apply this pipeline, the pixels $p_{i,j}$ need to be known in every image I_j. There are several ways to obtain these coordinates. A manual approach is to fit an ellipse to the mirror spheres as we explain in Section 4.3.4. This yields the sphere center in 2D as well as the 3D sphere position in camera coordinates. A second approach is to take a photo of the scene with a camera ring flash (*Canon MR-14EX TTL*) as proposed by Lensch *et al.* [Lensch03]. The flash will create a highlight on every sphere in the scene as shown in Figure 4.13. Each highlight is centered around the ray from the camera through the sphere center.

4.3.3 Calibration Setup

Next, we describe our capture setup as shown in Figure 4.12. The discussion includes the scene with the spheres, our metric floor mat, which is the basis for the ground truth measurements, and the camera we use.

Mirror Spheres: In this setup, we distributed the mirror spheres at arbitrary but known position on the floor mat. We use eight mirror spheres but only require a minimum of two spheres to calibrate the light source. Using more spheres naturally increases the robustness of the approaches. We evaluate in Section 4.3.5 to which extent fewer spheres degrade the accuracy of the results. Three of the eight spheres are placed at an elevated position on three stands that are 5 cm, 10 cm and 15 cm above

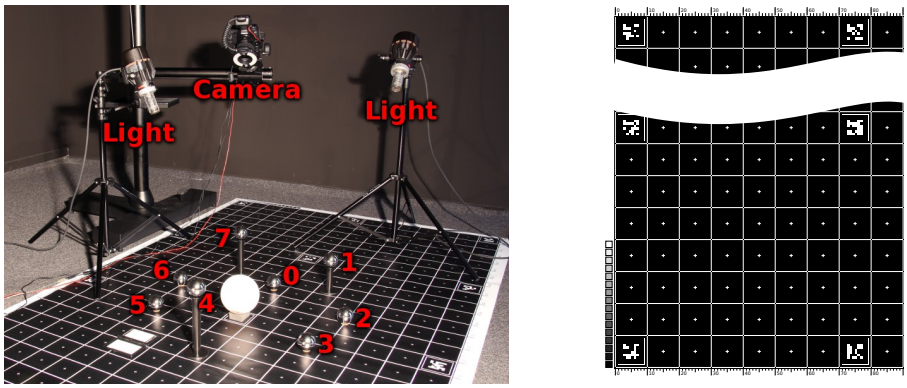


Figure 4.12: The capture setup. *Left*: The image shows the camera with ring flash attached, the light sources (we use only one at a time), and the spheres with corresponding numbers. *Right*: The 3 m × 1.6 m floor mat is the basis for our ground truth measurements.

ground. This avoids degenerate (planar) 3D point constellations in the Structure-from-Motion scene reconstruction described in Section 4.3.2. The quality of the spheres is quite relevant. We experimented with spheres of varying grade, and even slight geometric inaccuracies on the surface can lead to highlights that are offset by several pixels and markedly influence the stability of the results. We use high quality bearing balls with a diameter of 6 cm.

Metric Canvas: In order to obtain ground truth positions for both the spheres and the light source, a calibration target with metric information has been printed on a large canvas. Figure 4.12 (right) shows a cutout of the pattern. We used this canvas as floor mat and carefully placed the spheres at marked positions. The ground truth light positions have been measured using a plummet from the center of the light bulb to the floor mat. We expect that the accuracy of these measurements is in the order of millimeters for both the spheres and the light. This seems sufficient as the errors of the light estimation are orders of magnitudes larger.

Camera: We captured all photos using a *Canon 5D Mark II* camera with a *Canon EF 35mm F1.4L* prime lens. The intrinsic parameters of camera and lens have been calibrated prior to the evaluation using a checker board as described in Section 4.1. The calibration determines the exact focal length (we kept the focus point fixed for all photos), the principal point, and the radial distortion parameters.

Light: We used a *K5600 Joker-Bug 800* HMI lamp which produces a high light output by exciting a pressurized mercury vapor in the bulb. This lamp is particularly well suited for our task because it provides a good approximation of a point light source.

4.3.4 Preprocessing

In a preprocessing step, we first determine the distance d of each sphere from the camera center and the projection p of the sphere center onto the image plane. The sphere will project as an ellipse with parameters directly computable from the known

Figure 4.13: *Left*: We manually adjust the distance of a sphere until its projected ellipse fits to the image. The highlight was created with a ring flash and indicates the sphere center. *Right*: The bright light source creates highlights that are easy to detect automatically.

camera intrinsics [Hartley06] and the radius of the sphere. We manually adjust p and d until the rendered ellipse matches the image of the sphere as shown in Figure 4.13. This procedure could be automated by first segmenting the sphere, fitting an ellipse, and then recovering p and d as proposed by Wong *et al.* [Wong08]. However, manual parameter fitting seems appropriate for two reasons: Firstly, segmenting the mirror spheres is a hard problem due to low contrast between the spheres and the background. Secondly, we found manual parameter selection to yield higher accuracy—in the order of at most a pixel for p and a few millimeters for d.

We also run an automatic highlight detection that reliably selects the point of the light reflection on each sphere with subpixel accuracy. A simple but reliable procedure is to first apply a non-maximum suppression on the intensity image with a large radius. Then, for each maximum, we collect all pixels with an intensity value of at least $t < 1$ times the maximum intensity in a small radius around its position. The average coordinate of these pixels yields the final highlight location. We use HDR images and $t = 0.5$.

4.3.5 Evaluation

We first evaluate the techniques we introduced in Section 4.3.2, namely the *forward calibration* and the *backward calibration*. We do this for both varying camera positions and different light source locations. Afterward, we analyze how the number of spheres influences the calibration. This aspect is typically disregarded in other works which assume a fixed number of spheres. Finally, we evaluate the direct light source triangulation.

Dependency on Reflection Geometry

For a fixed set of spheres, the reflection geometry (see Figure 4.11) depends only on the relative positions of the camera and light source. We investigate the robustness of the forward and backward calibration with respect to varying constellations of those. We first captured two datasets with eight images from varying view points each. The ground truth light positions L_1, L_2 for the two datasets are as follows:

$$L_1 = (102.6, 0.0, 114.5), \qquad L_2 = (55, -35, 74.5).$$

After calibration, the light position is given in the local coordinate system of the camera. To study the variance and to compare against the measured ground truth,

Evaluation of Forward and Backward Calibration to Ground Truth [cm]				
Calibration and Dataset	Standard Deviation	RMS Distance	Min Error	Max Error
L_1 (fwd)	(2.8, 1.0, 2.9)	7.1	1.5	13.0
L_1 (bwd)	(1.2, 1.2, 2.0)	6.0	3.1	8.2
L_2 (fwd)	(1.1, 1.2, 1.5)	3.4	1.3	4.6
L_2 (bwd)	(1.1, 1.0, 1.3)	2.8	0.9	4.1

Table 4.1: Evaluation results for forward and backward calibration on two datasets. Light positions were estimated in all camera frames. We show the standard deviation of the light position and the RMS distance to the ground truth position, as well as the minimum and maximum error.

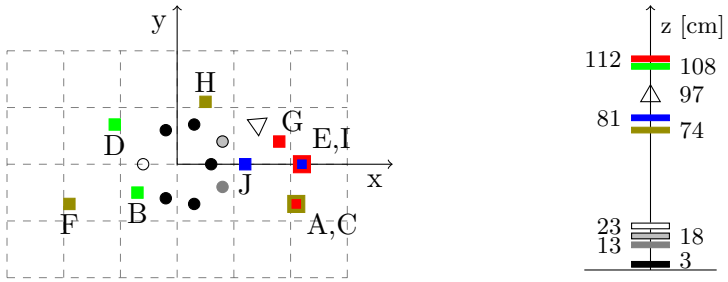


Figure 4.14: *Left:* Positions of spheres (circles), light sources (squares), and the camera (triangle) in the xy-plane. The grid lines are spaced 50 cm apart. Note that at some positions ($A+C$ and $E+I$) only the z-coordinate changes (the bigger square always corresponds to the light mentioned first) and that three spheres are placed at elevated positions. *Right:* Color coding of the z-component with height in cm.

we have to transform the light position into a global coordinate system. To do this, we determine a rigid, least squares optimal transformation [Umeyama91] from the estimated 3D sphere positions to the ground truth sphere positions derived from our metric canvas. Of course, this transformation will also include a small alignment error.

Table 4.1 shows the evaluation results for the forward and backward calibration. In particular, we computed the standard deviation of estimated light positions for all eight view points. We also computed the root means square (RMS) distance to the ground truth light position. Both methods, the forward and the backward calibration, perform similarly. We also notice that the error in dataset L_1 is larger than for L_2. The distance between light and spheres is about 1.5 m for L_1 and 1.0 m for L_2, so the larger error is plausible.

After varying the view point, we captured an additional dataset and moved the light source to 10 different positions while keeping the camera fixed. Figure 4.14 gives an overview over the sampled light positions and mirror sphere centers in the xy-plane together with a color-coding of the z-component.

The table in Figure 4.15 lists the distances—in centimeters—of our estimates compared to the ground truth position as error. It also contains the distance of the ground truth position to the center of the scene coordinate system. The plot next to it vi-

ID	Distance [cm]	Error fwd	Error bwd
A	133.4	4.2	2.8
B	116.0	6.4	6.3
C	157.2	6.8	4.6
D	125.9	7.7	8.7
E	156.8	8.4	7.4
F	125.4	10.6	8.9
G	144.8	7.9	7.3
H	95.5	4.2	4.8
I	136.3	5.7	3.4
J	100.4	3.6	2.9

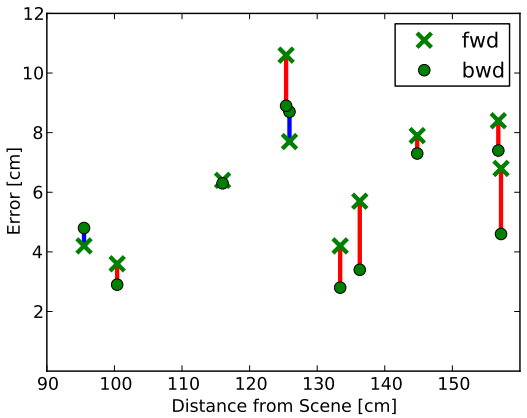


Figure 4.15: *Left:* Distance between ground truth and our estimates for varying light source positions according to Figure 4.14. *Right:* The error in general increases with distance. Light positions where the backward calibration outperforms the forward calibration are shown with red connecting lines. Blue indicates the opposite.

sualizes the same data and shows in red all cases where the backward calibration outperformed the forward calibration. It can be observed that the backward calibration yields slightly better results for most light positions, however, the difference between the approaches is small. The results also suggest that the benefits are greater for positions that are farther away. Overall, the accuracy is in the order of several centimeter which is only slightly better than the results of Powell *et al.* [Powell01] for the forward calibration.

Dependency on the Number of Spheres

The impact of the number of spheres on the robustness is rarely considered in light source estimation. The goal is rather to place as few calibration objects in the scene as possible. While our technique requires only a minimum of two spheres, the results so far have been computed with eight spheres. We believe that in many image-based reconstruction setups increased robustness is well worth the effort.

We run the proposed algorithm for all possible combinations of n out of 8 spheres. To reduce the number of possible combinations, we perform this evaluation only on light position E from the table in Figure 4.15. This position promises a challenging configuration because it has a relatively large error even for $n = 8$, and the light position has the second largest distance to the spheres. The standard deviation for fixed n in the second column of Table 4.2 gives an indication of the stability with respect to different sphere configurations and baselines. As expected, we see a strong decrease with growing number of spheres for both techniques.

The RMS distance of all $\binom{8}{n}$ results to the measured ground truth is summarized in the third column of Table 4.2 for $n = 2, \ldots, 7$. We observe that the error for both calibration methods decreases with an increasing number of spheres, see also Figure 4.16 which illustrates this for all light positions. Interestingly, the largest change occurs already at the transition from two to three spheres. We assume that a lot of other techniques which consider only two spheres could be improved just by

Number of Spheres	Standard Deviation [cm]	RMS Distance to Ground Truth [cm]
2 (fwd)	$(5.0, 1.5, 5.8)$	11.0
2 (bwd)	$(4.9, 1.6, 5.9)$	11.0
3 (fwd)	$(2.2, 0.7, 2.6)$	8.8
3 (bwd)	$(2.0, 0.7, 2.5)$	8.3
4 (fwd)	$(1.3, 0.4, 1.5)$	8.6
4 (bwd)	$(1.5, 0.5, 1.9)$	8.0
5 (fwd)	$(0.9, 0.3, 1.1)$	8.5
5 (bwd)	$(1.2, 0.4, 1.5)$	7.8
6 (fwd)	$(0.7, 0.2, 0.7)$	8.5
6 (bwd)	$(0.9, 0.3, 1.2)$	7.7
7 (fwd)	$(0.4, 0.1, 0.5)$	8.4
7 (bwd)	$(0.6, 0.2, 0.8)$	7.5

Table 4.2: Different number of spheres $n = 2, \ldots, 7$ used for calibrating light position E. Statistics in each row are computed over all $\binom{8}{n}$ possible combinations of spheres and evaluated for both techniques.

adding a third target.

So far, we have analyzed the combined error over all possible combinations. It is also interesting to ask whether some constellations might be better suited for reconstruction than others. We therefore take a more detailed look at the distribution of errors for all pairs which corresponds to all combinations with $n = 2$. This number is used in the majority of related approaches [Nayar89a, Powell01, Takai09], but those do not consider the impact of sphere placement. Figure 4.17 plots the distance of the reconstruction using individual pairs of spheres and the ground truth light position E. For 88 % of the pairs, the error is lower than 15 cm, and only three combinations lead to larger deviations.

These results were obtained for a fixed light source, but the error can also depend on the position relative to the pair of spheres. To keep the amount of possibilities tractable, we study this effect for all pairs that contain sphere number 0. Figure 4.18 shows the positional error for these pairs and all light positions. The values at position E thus correspond to the first row of the plot in Figure 4.17. Again, most of the errors are below 15 cm. Summarizing the observations from Figure 4.17 and Figure 4.18, we find that none of the combinations clearly outperforms the others for all light positions. Thus, we cannot detect a preferred arrangement of the spheres in the scene.

Pose Estimation and Direct Triangulation

Many image-based reconstruction tasks require to observe the target object from multiple camera positions, such as multi-view photometric stereo [Beljan12]. To estimate the camera pose either tracking markers have to be placed in the scene or features on the object need to be detected. If light estimation with mirror spheres is performed in such a context, the spheres can directly be used for pose estimation with the help

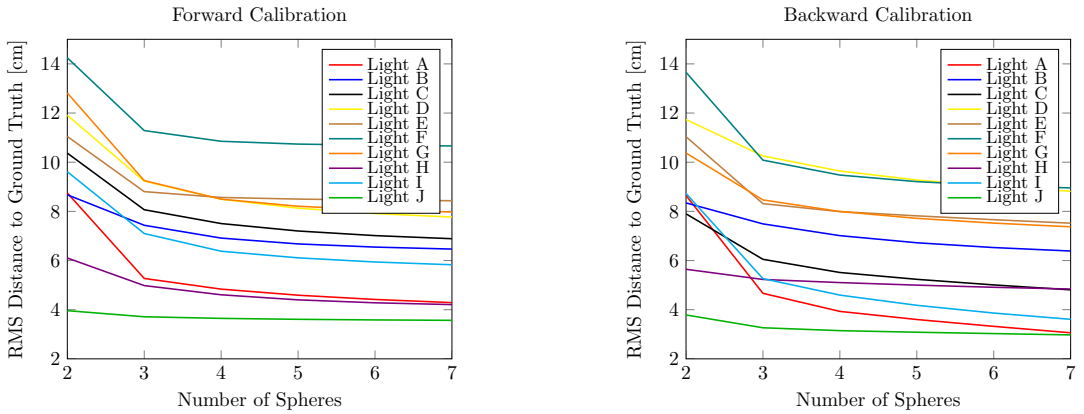


Figure 4.16: The forward and backward error for different number of spheres for all light positions. Each RMS error is computed over all combinations of n spheres out of 8.

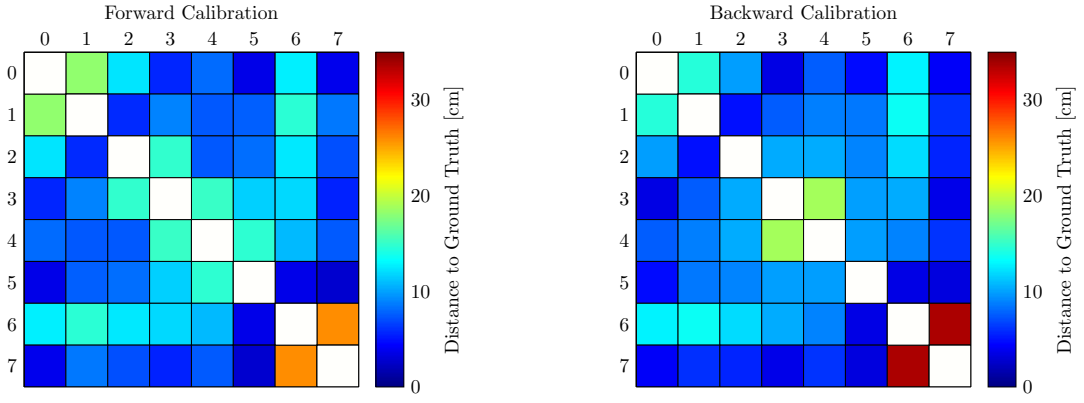


Figure 4.17: The forward and backward error for any combination of two spheres illuminated from position E.

of a ring flash, see Section 4.3.2. Additionally, a highly accurate light position can be recovered if at least two images show the light source directly as in Figure 4.19.

For each of the two initial datasets L_1 and L_2, we took additional images that contained both the spheres and the light source itself. Because the light source is extremely bright, we used the *B+W Gray Filter 72mm 110 E 1000x* which reduces the incoming light intensity by about 10 f-stops. This yields an extremely well localized point light in the image. The light can be automatically detected with subpixel accuracy using the same technique as described in Section 4.3.4.

The results for triangulating the light position and optimizing according to Equation (4.36) are given in the table in Figure 4.19. As can be seen, for the first dataset, the positions are highly accurate with errors of less than a centimeter although the distances between the cameras and the light source were about 4.5 m. For the second dataset, the distances between light source and the cameras were about 2.5 m, and the positional error is in the order of the uncertainty of the ground truth measurements. In terms of accuracy, the direct approach is clearly superior to any of the reflection-based calibrations.

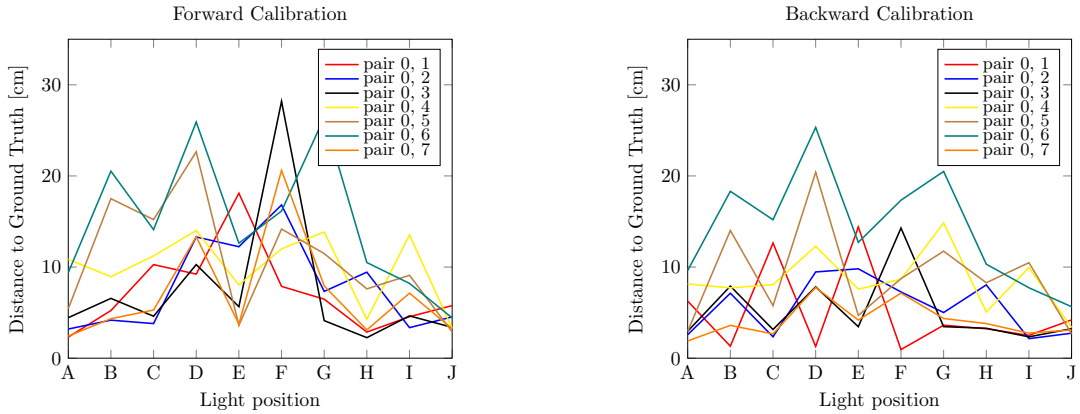
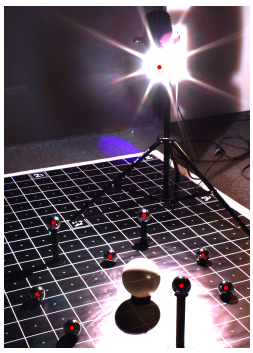


Figure 4.18: Error for all light positions and all pairs of spheres that contain sphere number 0.



Evaluation of Direct Light Triangulation [cm]		
Ground Truth	Estimate	Error
$(102.6, 0.0, 114.5)$	$(102.4, -0.2, 113.7)$	0.9
$(55.0, -35.0, 74.5)$	$(55.0, -35.1, 74.3)$	0.2

Figure 4.19: *Left:* Tone mapped example picture for direct light calibration with sphere centers and detected light position marked in red. *Right:* Evaluation of light source positions obtained through direct triangulation followed by bundle adjustment. The result is within millimeters from the measured ground truth position.

4.4 Calibrated Photometric Stereo

Before we delve into more challenging input data, it is worth to first study the errors that might arise in a controlled setup. This will provide an intuition of what we can expect at best in an uncontrolled setting. The experimental comparisons in related works are either performed with hardware that does not resemble state of the art consumer cameras [Silver80, Ray83, Tagare91, Sato95], do not contain a quantitative evaluation on real images [Wu06, Wu10, Alldrin08, Verbiest08, Yoshiyasu11, Ikehata12, Shi12b, Shi12a, Yu13, Park13], or consider a special setup and technique, *e.g.* [Tunwattanapong13]. Others compare the geometry error of a final integrated surface [Silver80, Georghiades03, Simakov03, Zhou13]. It is also common to compare a more advanced technique against standard photometric stereo [Tan07, Shi10, Sunkavalli10, Higo10, Tan11, Papadhimitri13]. In our experiments, we would like to study a straightforward photometric technique in a controlled but simple capture scenario with a digital consumer camera.

Calibrated photometric stereo techniques have two inputs: the image intensities and the light directions. In any experimental setup, these "observables" will exhibit

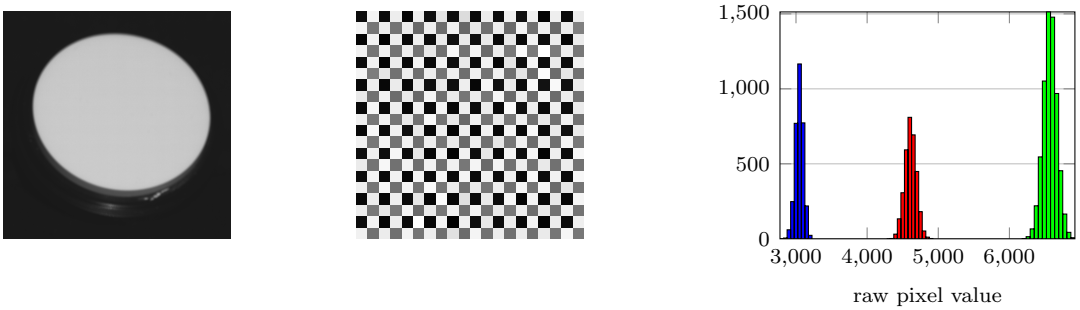


Figure 4.20: Raw sensor values. *Left*: Raw image of the diffuse reflectance standard. *Middle*: A closeup of the unprocessed sensor values shows the color filter array. *Right*: Histogram over 16×16 pixels in 51 images for the red, green, and blue channels.

inaccuracies due to the measurement process. Those are, however, not the only sources of error that impact the reconstruction of normals n_j. The assumptions required for Equation (2.27) to hold might also be violated. These effects, which are not modeled by the theory, are difficult to detect and quantify. In practice, it is usually not clear how to distinguish between measurement noise and model violations. Accordingly, overall errors will include a combination of both sources. We will discuss some error sources that come to mind and give an intuition about their magnitude through simple experiments.

4.4.1 Model Assumptions and Error Sources

The typical formulation

$$I_{i,j} = \rho_j \max(\langle n_j,\ D_i \rangle, 0) \tag{4.37}$$

requires a distant point light source of either constant or known luminance. It does not include interreflections and cast shadows. Furthermore, it assumes a perfect Lambertian reflection and linear camera response. We combine the sensor sensitivity, light source luminance, and per-pixel albedo in a single term ρ_j.

Sensor Noise

Even if the experiment fulfilled all assumptions of the model, the intensities $I_{i,j}$ would still be subject to measurement noise. To study its distribution, we analyze 51 images of a planar reflectance standard under constant illumination. All experiments are conducted with a Canon EOS 700D consumer camera which has a spatial resolution of 5208×3476 pixels. We always use a low sensitivity setting (ISO 100).

Most digital cameras use a sensor with an added color filter array. Thus, individual sensor elements capture contributions either of the red, green, or blue channel. These contributions have to be interpolated (*demosaicked*) over the image plane to produce a full-sized, colored image with RGB information at each pixel. We circumvent this interpolation step and operate on raw sensor output as shown in Figure 4.20. This Bayer pattern [Bayer76] leads to 25 % of the pixels having blue or red characteristics and 50 % available for green.

Experiment	red			green			blue		
	μ	σ	μ/σ^2	μ	σ	μ/σ^2	μ	σ	μ/σ^2
planar (1/20 s)	2236	49	0.93	3202	57	0.99	1481	35	1.21
planar (1/10 s)	4585	81	0.70	6544	96	0.71	3022	55	1.0
planar (1/8 s)	5809	101	0.57	8261	117	0.60	3827	67	0.85
sphere (1/10 s)	2299	52	0.85	3274	68	0.71	1501	43	0.81
sphere (1/10 s)	2994	72	0.58	4287	86	0.58	1990	46	0.94
sphere (1/10 s)	3882	70	0.79	5557	95	0.62	2566	54	0.88

Table 4.3: Statistics of intensity samples under constant illumination (51 images). *Top rows:* A planar target evaluated in a 16×16 window. *Bottom rows:* Three 10×10 windows on a sphere corresponding to Figure 4.21 (left).

Even if no light reaches the sensor, the camera produces a nonzero output. We take several pictures with the lens cap attached and find that this effect amounts to 2050 intensity levels. We experimentally found this to be independent of the color channel and pretty stable for varying exposure times. We subtract this value from all sensor readings. From over-exposed pictures, we also find the saturation threshold of 13 583 units.

We select a 16×16 rectangular region and compute histograms over all 51 images for each of the pixel classes. Figure 4.20 shows how the values are distributed. The standard deviations $\sigma_R = 81, \sigma_G = 96, \sigma_B = 55$ correspond to 0.7 %, 0.8 % and 0.5 % of the 11 533 bins. If we had instead used the linear but color processed images, this would have increased to 1.1 %, 0.8 % and 0.9 % respectively.

To transfer these findings to other experiments, we have to ask whether the deviations would change for different shutter speeds. The top three rows of Table 4.3 show the results of the same testing procedure applied to three exposure settings. We notice a clear dependence of the standard deviation on integration time, and thus on μ. The ratio of μ and σ^2 is, however, not exactly constant—neither between channels nor between exposures. Thus, the deviations are not caused purely by "shot noise", which is Poisson distributed. As we would expect from a linear sensor, the mean intensity is almost proportional to exposure time. The deviations might be attributed to inaccurately reported shutter speeds.

Next, we exclude any effects related to exposure time and keep it fixed at 1/10 s. To achieve different mean intensities, we either have to change the illumination or use a non-planar target. We choose the latter and take another 51 images of a diffuse sphere. This means that each pixel now observes a surface point with slightly differing normal. We therefore reduce the size of the window to 10×10 pixels. Figure 4.21 shows the three windows in different regions of the sphere corresponding to high, medium, and low intensity values. Again, Table 4.3 does not indicate a straightforward relationship of mean and standard deviation that could be parametrized easily.

We conclude that there is a slight benefit in working on raw data and that we cannot derive a specific noise model. The *coefficient of variation* σ/μ computed from the samples ranges between 1.4 % and 2.9 % in our experiments.

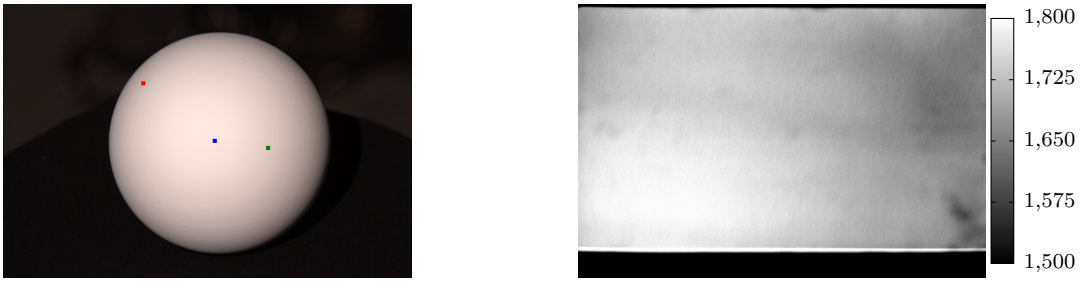


Figure 4.21: *Left:* Pixel windows (enlarged) used in sensor noise evaluation. *Right:* Intensity falloff on a planar gray card caused by light source distance (positioned to the left and front, *i.e.* the bottom left corner of the image) and vignetting. We show the raw sensor output of the green channel after black level subtraction.

Falloff

Another source of error is introduced by the optical system of the camera. So called *vignetting* leads to reduced image intensities in peripheral regions compared to the center. To assess whether this effect is relevant in our setting, we capture a flat gray card of 21 cm × 30 cm almost perpendicular to the optical axis (lens used: *Canon EF 70-200mm F2.8L* zoomed to 110 mm). The result in Figure 4.21 does not show the radial falloff typically associated with vignetting.

We observe, however, that the intensities are not quite constant over the whole area. First, there are some dark spots at the lower right which are due to a stain on the card. Second, the lower left is 200 to 300 sensor units brighter than the upper right. This effect is probably caused by the $1/r^2$ falloff in irradiance since the light source shines from that direction. On the one hand, this shows that the light is not infinitely distant and violates our assumptions. On the other, the impact is rather small in smaller regions. The diffuse sphere has a diameter of 10 cm, and the falloff would only amount to about 100 units (0.9 %) in the worst case.

Linear Response

To verify that the raw sensor values behave linearly, we capture an exposure series of 19 images of the sphere from 1/640 s to 2.5 s. The curvature of the sphere ensures a good sampling of the full intensity range. From such an exposure stack, the response curve can be recovered very reliably using the method proposed by Robertson *et al.* [Robertson99]. The result in each color channel is plotted in Figure 4.22 and deviates only marginally from the best fitting line. Thus, we can assume the requirement of linear intensities as fulfilled.

Camera Motion

Accurate correspondences between pixels in all images are important for photometric stereo. We use a solid tripod and a remote trigger to reduce the sources of camera motion. Unfortunately, the camera does not allow to fix the mirror when shooting in remote control mode. That means that the mirror in front of the sensor has to be

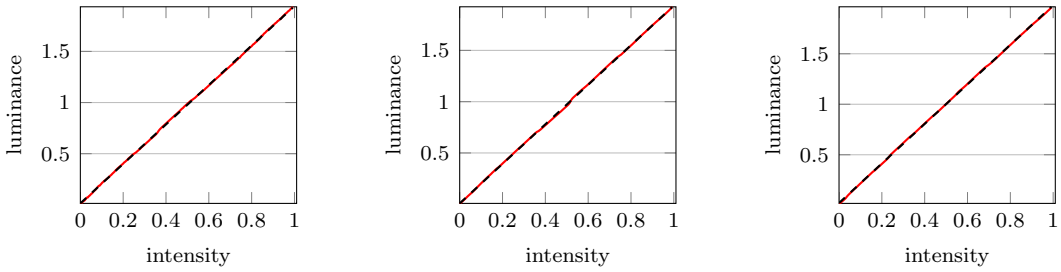


Figure 4.22: Linear sensor response. The estimated response curve (red) in each color channel (*from left to right*) deviates only marginally from the best linear fit (black).

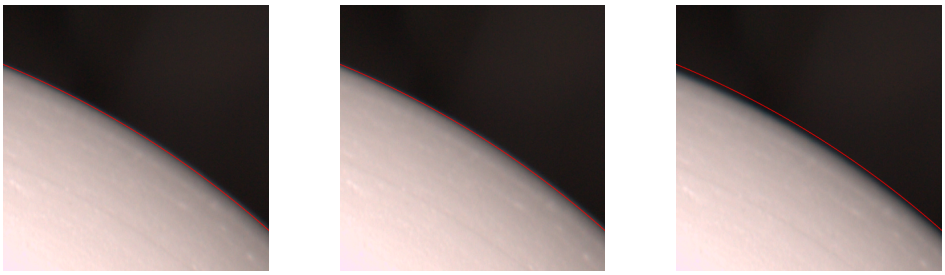


Figure 4.23: Camera shake in a 51 image sequence. We manually fit a circle to the diffuse sphere in the last image of the sequence (*left*). The same circle is drawn on the next-to-last (*middle*) and first (*right*) image. The deviations accumulate to an error of 4-5 pixel.

flipped away in every image, possibly introducing a shake.

We look at the impact of camera shake in a sequence of 51 images. First, we manually fit a circle to the observed boundary of the sphere in the last image. Then, we draw the same circle onto the next-to-last and the first image. Figure 4.23 shows that there is hardly any motion between the two consecutive frames. Comparing the error over the whole sequence, however, yields a deviation of 4 to 5 pixels. In relation to the whole image, this corresponds to about 0.1 %. We did not observe a distinct pattern in the occurrences of these shifts.

Such an error will have less implications in the center region of the sphere. There, points that are 5 pixels apart still observe almost the same normal. Towards the boundary, however, we have to expect a decrease in performance. It is important to keep this in mind when evaluating later experiments. There is hardly anything we can do about these errors in our setup.

Light Directions

We have discussed in Section 4.3 that calibrating the position of a near point light source is prone to errors in the range of centimeters. Now, we place the light source much farther (3.5 m) away from the scene to approximate the requirements of calibrated photometric stereo. We again rely on mirror spheres to estimate the light source direction and use the same procedure to automatically obtain highlight positions. Under the assumption of an orthographic camera, the sphere center and radius are found by manually fitting a circle in the image. From the known sphere center,

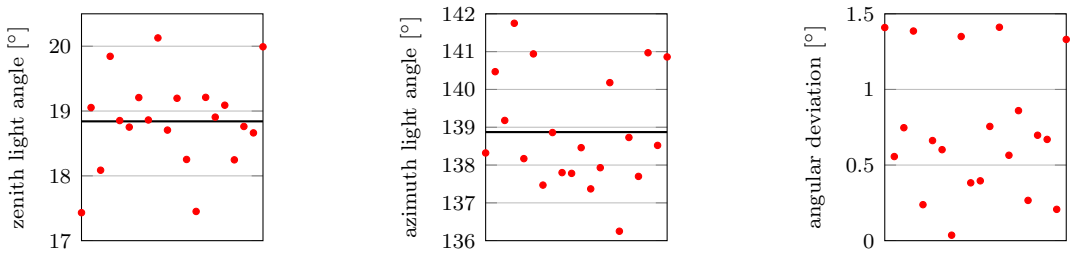


Figure 4.24: Varying estimates for fixed light direction. A mirror sphere is placed at 20 positions in the scene to reconstruct the direction of the light. The deviations in recovered zenith (*left*) and azimuth (*middle*) angle are shown with their respective means as black line (computed with the smallest and largest value left out). The angular deviation from the mean direction stays below 1.5° (*right*).

highlight position, and respective normal, we compute the light direction in camera coordinates.

To obtain an estimate of the error, we keep the light fixed and move a sphere to 20 different positions in the scene. We keep these within a radius of about 20 cm from the location of the diffuse sphere in later experiments. For each position, we compute the azimuth ϕ and zenith θ angles of the light direction with respect to the camera coordinate system (the z axis is aligned with the optical axis). The results are plotted in Figure 4.24 and show that the estimation of the zenith angle is more accurate than for the azimuth—at least for this light configuration. We also assess the spread of the computed light directions. The angular deviation from the mean direction is below 1° in most cases.

Note that this error again contains all possible sources that we cannot distinguish such as detection errors, not-modeled camera distortions, not truly directional lighting, imperfect spheres, *etc*.

Further Deviations

Here, we identify further deviations from the ideal Lambertian assumption that commonly occur in practice. We also explain how we counteract them in our experiments.

Specular Reflection and Highlights: We use objects made of *Spectralon*, which is a special material used for optics applications. Its porous subsurface structure makes it a highly Lambertian reflector.

Shadows, Interreflections, and Ambient Light: We use a sphere and cover all nearby surfaces with black cloth to minimize interreflections. There are no cast shadows on the sphere because it is convex. Attached shadows can be omitted based on simple thresholding. We conduct the experiment in a large room painted in black to minimize ambient light.

Non-parallel Light and Camera Rays: We place the light 3.5 m away from the target sphere, which has a radius of 5 cm. While this does not ensure a truly distant illumination—as indicated by the previously discussed falloff—it is a reasonable

approximation. The camera is equipped with a 135 mm telephoto lens and placed 2.5 m from the object. We validate this approximation by looking at the silhouettes of spheres and find that they are almost perfectly circular.

Light Variations: Variations in light intensity and spectral output between images or between surface points are another source of error. We use a single metal-halide lamp and take care to place it at the same distance of the scene—checked with a tape measure—in each image to reduce intensity variations. We also throw away images whose brightest intensities differ throughout the series. After a warm-up phase, the lamp has a daylight spectrum which should be rather stable during a single capture session.

Others: We deem all other effects, such as spatially non-uniform sensors and/or optics (*e.g.* aberrations), optical properties of the medium, limited spatial resolution of the sensor, blurring of incoming light due to finite apertures, *etc.*, less significant.

4.4.2 Experiments

Now that we have an intuition about the different sources of error, we can conduct actual reconstruction experiments. We use the same lamp and camera as described in the previous section. The diffuse sphere is our target object. To estimate the light direction, we place two mirror spheres in the scene and average their respective light vectors.

For each light position, we first take a picture that shows the mirror spheres and highlights. Next, we cover them with black cloth to eliminate reflections onto the target. We then take ten pictures of the diffuse sphere and average them in order to reduce noise. We also threshold each color channel to define a mask of non-shadowed pixels. We repeat this procedure for nine different light directions.

Before we try to invert Equation (2.27), we should verify how well reality fits the model. This is not trivial because of the various sources of error that are not covered, but we can at least make an approximate check. In the Lambertian case, the intensity matrix should be close to a matrix of rank three. In our experiment, the first three singular values range from 6 to 110, whereas the remaining ones are below 0.5. The Frobenius norm of the best rank three approximation is 0.72. These values are not easy to interpret, but they indicate that the model is at least a reasonable approximation.

Another way of verification is based on our assumption of a constant luminance exiting the light source. If this is the case, the brightest pixels in each image should have the same intensity—even though their location within the image changes. Figure 4.25 shows the sensor values of the 1000 brightest pixels in the blue channel of each image. They are all within a small range of about 75 units, which lets us assume that the Lambertian model is fulfilled quite well. We also notice that three light positions clearly deviate more than the others, which is probably caused by an offset in light distance. We therefore exclude those from further considerations.

A third strategy to assess how well the model suits our data is to make predictions and compare them against the observations, *i.e.* to perform *cross validation*. For each pixel, we reconstruct the scaled normal from three images and then predict its

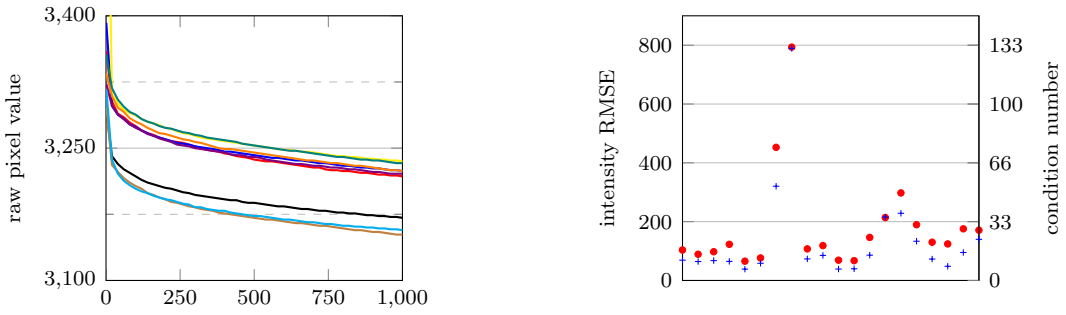


Figure 4.25: Testing the model assumptions. *Left:* The sensor values of the 1000 brightest pixels are plotted for each light position. In a perfectly Lambertian model, they would all coincide. Three images are obviously less bright and are thus excluded from further experiments. *Right:* We reconstruct normals for each subset of three out of six images and predict the intensities in the other three. The red dots show the intensity differences to the actual image observations squared and summed over all pixels. We also plot the condition numbers of the light matrix (blue) and observe that they are correlated with the intensity errors.

Work	Technique	Comparison Against	Angular Error
[Ray83]	standard PS	Lambertian sphere	$< 5°$
[Tagare91]	standard PS	Lambertian sphere	$4.8°$
[Abrams12]	outdoor, webcam, Sun	Google Earth	$20°$
[Tunwattanapong13]	special illumination	Lambertian sphere	$> 5°$
[Wu13]	dense PS	specular sphere	$4°$
[Shi10]	uncalibrated PS	standard PS	$6 - 7°$
[Favaro12]	uncalibrated PS	standard PS	$5 - 12°$
[Papadhimitri13]	uncalibrated PS	standard PS	$2 - 3°$

Table 4.4: Examples of "ground truth" choices and quantitative results in the literature.

intensity in all other images given the light directions. If the model is violated, these predictions deviate in general from the actual images. We use the root mean square error of all image intensities in the intersection of the masks as a measure. Figure 4.25 shows the result of running this test on all possible combinations of three out of six images. We observe that the error is mostly low apart from a few outliers caused by ill-conditioned lighting matrices.

Finally, we evaluate the quality of the reconstructed normals. The question is what we accept as a reference to base the evaluation on. Table 4.4 lists some examples of common choices in the literature. A frequently used approach is to compare one reconstruction technique against another one that is assumed to be more accurate. The other technique is, however, subject to the same error sources. Instead, one can compare against a reference object of known shape. This raises the question of how accurate the object really is and how well it can be aligned with the reconstruction. We choose the second approach and use an ideal sphere as ground truth. We do, however, not have any guarantees about the accuracy of the surface of the Spectralon sphere.

Figure 4.26 shows the qualitative results as a normal map for all pixels that were never in shadow for any of the six light sources. The histograms over all pixels in

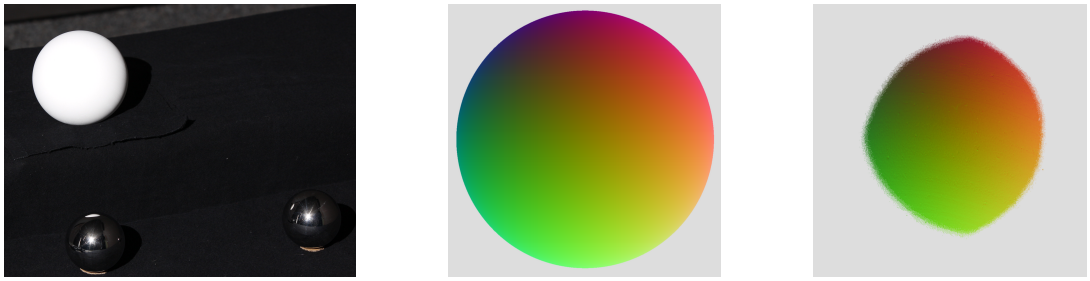


Figure 4.26: Qualitative results. *Left*: Cropped input image showing the target sphere and mirror balls. *Middle*: Ideal ground truth used for comparisons. *Right*: Pixels that are above the shadow threshold in all six images are used to reconstruct a normal map.

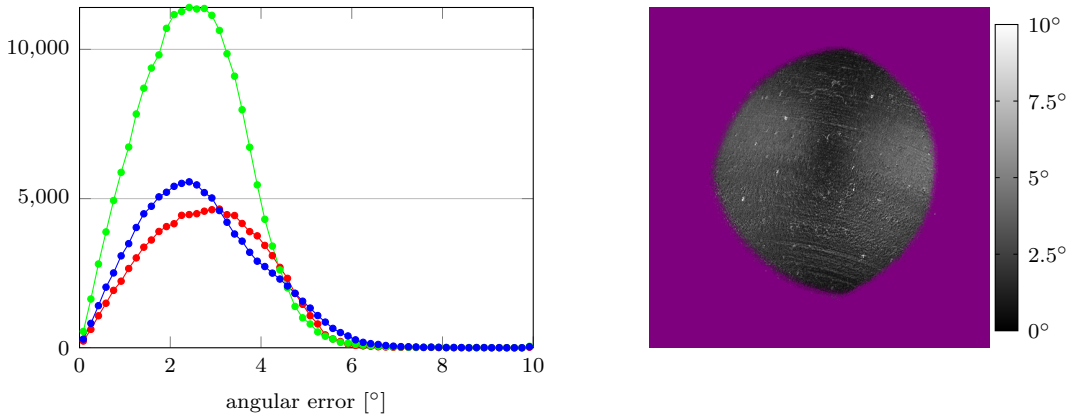


Figure 4.27: Quantitative results. *Left*: Histogram of angular deviations in all three color channels (the dominance of the green channel is caused by the Bayer pattern). *Right*: The spatial distribution of angular errors shows fine ridges and some dents present in the target sphere.

Figure 4.27 (left) provide a quantitative evaluation and show that the deviation from ground truth is about 2.5° in general. That is better than most of the results in Table 4.4 and gives an impression of what we can expect from a carefully calibrated setup. The error is spatially varying because the real world surface has fine ridges and several pronounced dents that deviate from a true sphere, see Figure 4.27 (right). We also notice deviations due to the structure of the porous material of the sphere.

4.4.3 Error Analysis

We have studied the experimental results of calibrated photometric stereo both qualitatively and quantitatively. The question is, however, whether we have achieved the best possible reconstruction given the data. To answer this question, we perform an error analysis.

One approach to this end is presented by Ray *et al.* [Ray83]. They assume a setup with three light sources on a circle around the optical axis. Then, the photometric stereo problem can be solved in closed form, *cf.* Equation (2.30), which we encode in

a function h:

$$n = h(I_1, I_2, I_3, \theta, \phi_1, \phi_2, \phi_3). \tag{4.38}$$

A first order Taylor expansion then yields an approximation of the error in gradient space $n = (-p, -q, 1)$ given the deviations $dI_1, dI_2, dI_3, d\theta, d\phi_1, d\phi_2, d\phi_3$ of the input data:

$$dp = \sum_i \frac{\partial h_1}{\partial I_1} dI_i + \sum_i \frac{\partial h_1}{\partial \phi_1} d\phi_i + \frac{\partial h_1}{\partial \theta} d\theta, \qquad dq = \sum_i \frac{\partial h_2}{\partial I_1} dI_i + \dots \tag{4.39}$$

In our case with six light sources, a solution h involves the pseudo inverse and computing the derivative is non-trivial. We therefore use a different strategy and formulate the image formation with random variables whose distributions we then sample. We assume that the light source power and the albedo are constant and equal to one. Given the ground truth normal n_{GT} and measured light directions $\hat{\Theta} = (\theta_1, \dots, \theta_6)$, $\hat{\Phi} = (\phi_1, \dots, \phi_6)$, we model the observables with normal distributions

$$I \sim \mathcal{N}(A(\hat{\Theta}, \hat{\Phi}) \cdot n_{\mathsf{GT}}, \sigma_I), \quad \Theta \sim \mathcal{N}(\hat{\Theta}, \sigma_\theta), \quad \Phi \sim \mathcal{N}(\hat{\Phi}, \sigma_\phi) \tag{4.40}$$

where the standard deviations $\sigma_I = 0.008, \sigma_\theta = 0.698°, \sigma_\phi = 1.451°$ correspond to the experimental derivation in Section 4.4. $A(\hat{\Theta}, \hat{\Phi})$ is the light matrix. Then, we draw random samples $\tilde{\Theta}, \tilde{\Phi}, \tilde{I}$ from each distribution and compute the resulting normal as

$$n = n(\theta_n, \phi_n) = \frac{A(\tilde{\Theta}, \tilde{\Phi})^\dagger \tilde{I}}{\|A(\tilde{\Theta}, \tilde{\Phi})^\dagger \tilde{I}\|} \tag{4.41}$$

where $A^\dagger$ denotes the pseudoinverse.

Figure 4.28 (left) shows the histogram of the angular error over all normals considered for the experiments, *cf*. Figure 4.26 (right), and over 10 000 iterations of random sampling. We observe that given the uncertainties in the input data, we cannot expect to obtain much better results than a 1° error. The histogram in Figure 4.27 (left, green channel only) is pretty similar and indicates that our experimental results are quite close to the best possible ones. To study the spatial distribution, Figure 4.28 (right) shows the average error over 10 000 iterations at each pixel. Again, the distribution is similar to the experiment in Figure 4.27 showing a general increase of error towards the left and right.

To inspect the differences between the reconstruction and the ground truth more closely, we also consider errors $(d\theta, d\phi) := (\theta_{n,\mathbf{GT}} - \theta_n, \phi_{n,\mathbf{GT}} - \phi_n)$ in the zenith and azimuth directions instead of the angular error $\operatorname{acos}\langle n,\ n_{\mathsf{GT}}\rangle$. Figure 4.29 (left) contains the joint histogram of $(d\theta, d\phi)$. We observe that deviations in azimuth are larger than in the zenith angle which is to be expected because the azimuth angles of the light source also have a larger uncertainty. In contrast to Figure 4.28 (left)—which is not peaked at zero—we also observe that the joint histogram is centered around $(0, 0)$. The discrepancy can be explained since several combinations of $d\theta, d\phi$ can lead to the same angular error.

It is insightful to study which of the input variables I, θ, ϕ contributes most to the overall error. Figure 4.29 (middle) demonstrates this by sampling only one of these random variables and using ground truth information for the other two. The

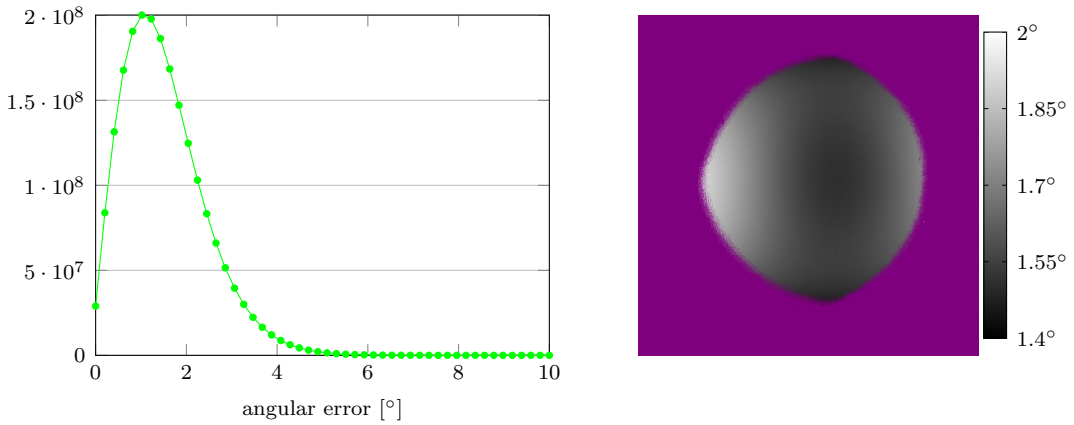


Figure 4.28: Simluation results. *Left*: Histogram of angular deviations in the green channel over all pixels and over 10^4 iterations of randomly sampling the input variables. *Right*: The spatial distribution of the angular error. At each pixel, the average error over 10^4 iterations is shown.

histogram indicates that in our setup the main contribution to the ensuing errors stems from the uncertainty in the intensities I. Another interesting question is how the number of light sources impacts the results. We therefore consider all possible combinations of $k = 3, 4, 5$ out of six light sources. The histograms over all these combinations for varying k are plotted in Figure 4.29 (right). As expected, we observe that increasing the number of light sources leads to smaller deviations in general.

Finally, we ask how the light constellation influences the stability, which we investigate for a simplified example. Assume that three lights are distributed evenly on a circle around the optical axis. We vary the common zenith angle $\theta_1 = \theta_2 = \theta_3 = \alpha$ from 5° to 60° and again simulate errors in the input variables with the same standard deviations as in the six light source setting. Figure 4.30 shows that for small α the errors are rather large which is due to the ill-conditioned light matrix. For larger angles, the histograms—computed only over pixels that are never in shadow—are shifted towards the left indicating smaller errors in general. This leads, however, to a much larger percentage of shadowed pixels on the sphere and less complete reconstructions as illustrated by Figure 4.30 (right). Finding the optimal trade-off depends on the use case and is left for future work. A starting point is provided by Drbohlav and Chantler [Drbohlav05] who, however, ignore errors in the light source directions.

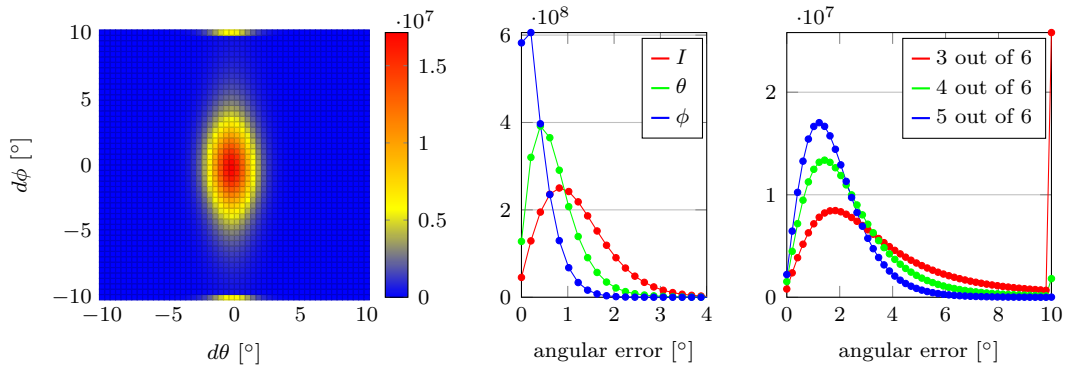


Figure 4.29: Error analysis. *Left*: The joint histogram of deviations in θ_n, ϕ_n computed over all pixels and 10^4 sampling iterations. The border bins collect all deviations beyond the displayed range of $(-10, 10) \times (-10, 10)$. *Middle*: Histogram of the angular error if variation is only simulated in either I, θ, or ϕ. *Right*: Influence of different numbers of light sources used for reconstruction. For each $k = 3, 4, 5$, the histogram—divided by $\binom{6}{k}$ for normalization—is computed over all pixels and all possible combinations of k out of six light sources. The number of sampling iterations per light configuration is 10^3.

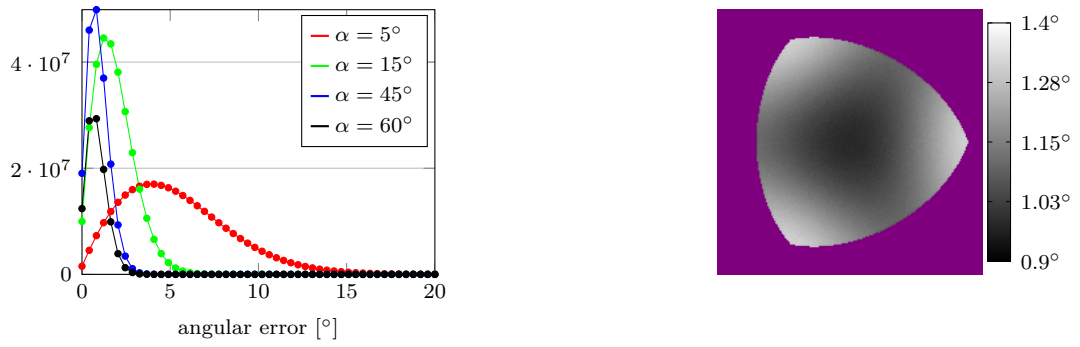


Figure 4.30: Simulations for three sources on a circle around the optical axis ($\phi_1 = 0°$, $\phi_2 = 120°$, $\phi_3 = 240°$, $\theta_1 = \theta_2 = \theta_3 = \alpha$). *Left*: Histogram of angular deviations over all pixels and over 10^4 iterations for varying α. Pixels that are in shadow from at least one source are discarded and lead to decreased coverage for larger α (5°: 99.6 %, 15°: 95.3 %, 45°: 59.9 %, 60°: 34.1 %). *Right*: The average error at each pixel over 10^4 iterations for $\alpha = 45°$.

4.5 Discussion

This chapter introduced several aspects of calibration in the context of appearance reconstruction. We discussed about standards that describe the radiometric calibration and combined them to recover per-pixel absolute luminance on Internet images. We also touched on the topic of "perceptually accurate" reconstruction using the example of tone mapped images captured under scotopic conditions. Then, we considered a controlled environment and presented several techniques to estimate the position of a point light source. Finally, we had a look at calibration with respect to photometric stereo and studied various sources of error experimentally.

The idea to use the Moon as a calibration target to assess the reliability of luminance values computed from metadata was only partially successful. The big impact of the atmosphere on incoming luminance prevented us from a more rigorous quantitative evaluation. Qualitatively, the radiometric model is sufficiently accurate to provide a basis for perceptual tone mapping algorithms. This enabled us to predict the low light impression of a human observer more plausibly than the original image does. This example shows that it is important to consider absolute luminances if we think about images—or reconstructions based on them—as means to fully capture our perception of the visual world.

For the calibration of close-by point light sources, we have found that both forward and backward calibration are accurate in the order of centimeters. Inspired by the performance of the image space error in bundle adjustment (*cf*. Section 4.1), we originally expected a more obvious advantage of the novel backward calibration. Nevertheless, our experiments show that the proposed technique leads to smaller errors in general, see Figure 4.15, and should be preferred. The evaluation also indicates that the spatial arrangement of the spheres can have a considerable impact on the performance in both cases. This aspect has been ignored in other works without evaluating whether it influences the results. We could, however, not make out a constellation that is preferable for all light source positions.

Any technique based on reflections at spherical surfaces is faced by the challenge that small uncertainties in detected highlights or estimated sphere positions are amplified when estimating the light source. Therefore, the second approach that we presented achieves a much higher accuracy by directly observing the light source in multiple images. However, it requires more images, and the camera has to be registered with the target scene, which can introduce an additional error. An alternative might be to combine the reflection based approaches with intensity measurements, *e.g*. on a diffuse sphere, to better constrain the distance of the light source.

Section 4.4 provides a reference for the overall quality of photometric stereo that we can expect. We found that the best results under carefully calibrated conditions deviated by $\approx 2.5°$ from the ground truth normals. This is very close to the theoretical limit of $1°$ for our setup which we determined through simulation of the uncertainty in the input variables. Approaches that operate on uncontrolled data will certainly lead to larger errors. The interesting question will thus be "How close can they come to this result?".

Chapter 5

Photometric Stereo for Outdoor Webcams

In Chapter 3, we have seen the progress that has been made on photometric stereo techniques. We have also discussed in Chapter 4 that calibrating a photometric setup is a crucial but non-trivial preprocessing step. Both aspects, the calibration and the reconstruction, are usually only considered for controlled conditions such as an optics laboratory. In this chapter, we are pushing the limits of photometric stereo algorithms to gain insights and a better understanding of what is possible in an uncontrolled setting.

A move towards less controlled input data can be observed in several disciplines of computer vision. There is a general consensus that one of the most uncontrolled and challenging data sources are online image collections. These have emerged as a valuable source for various applications including multi-view stereo [Furukawa10a, Goesele07], segmentation [Simon08], or reflectance recovery [Haber09]. However, they have not yet been considered by the photometric stereo community since they do not fulfill one of the fundamental assumptions in this field: *known pixel correspondences between images*. As a first step, our idea is therefore to focus on image sequences from publicly available webcams on the Internet. These images have an inherent correspondence since they are all taken from the same viewpoint. On the other hand, this kind of data is still completely uncontrolled and provides similar challenges and opportunities to image collections—apart from, for example, assuming a constant camera (hardware) in all images.

Thus, one of the motivating questions in this chapter is "Is it possible to apply photometric stereo on webcam images?". Conceptually, this should be easy, *e.g.*, by combining a robust photometric stereo approach with an illumination estimation technique. However, naively applying existing techniques to such datasets leads to poor results because some unique problems have to be addressed. We either have to transfer lab-based methods for camera and lighting calibration to uncontrolled settings or find new ways specific to Internet webcams.

We will pursue the latter and present a complete calibration pipeline. Without access to the scene, the key insight is to find more general concepts that are present no matter the input data and exploit them for calibration purposes, *e.g.* symmetries, specific patterns, or other constraints. We found such an aspect by restricting our approach to outdoor scenes. This gives us some prior knowledge about the scene:

the illumination is mainly a combination of the Sun and the sky light and varies not randomly—disregarding clouds or weather influences—but according to certain patterns given by the rotation of the Earth around its axis and around the Sun. We can then exploit this knowledge to calibrate webcams radiometrically and geometrically. Furthermore, this provides a partial calibration of the illumination since the direction of the sunlight can be derived from the time of day and astronomical data, *cf*. [Reda04].

A second challenge lies in the scene content itself. The objects can exhibit arbitrary reflectance properties and need not follow a Lambertian assumption. Furthermore, surface reflectance will be spatially varying, *e.g*. the wall of a house differs from its roof. As discussed in Chapter 3, only few photometric stereo techniques are sufficiently robust and general in their assumptions to cope with such data. For example, the one other approach shown to operate on webcam images, [Abrams12], recovers only diffuse surfaces. We therefore develop an image creation model for outdoor scenes that is based on a small set of basis materials which are then mixed at each pixel. This allows for spatially-varying, non-Lambertian reflectance.

Another challenge is the proper selection of input images. Webcams typically generate novel images every few seconds or minutes. This leads to a huge amount of data that has to be reduced for practical reasons, such as computational resources, but also because many images have to be considered as outliers, *e.g*. night-time scenes. A robust view selection is the crucial part of multi-view stereo algorithms for Internet data, see *e.g*. [Frahm10], but it has not been considered for requirements as posed by photometric approaches. In our case, this is the conformance with an image formation model which assumes a clear sky with a bright Sun as the primary light source. We propose an automatic selection scheme that analyzes the sky and object regions of the webcam images. It also takes the two-dimensional variation of solar position into account to ensure that normals can be recovered unambiguously.

Accounting for the huge variations possible in such general input data leads to a rather complex model $\mathcal{M}$ with several parameters. Among these are the normal map N which we want to reconstruct together with reflectance properties ρ. As in many inverse vision problems, we formulate the reconstruction task as a minimization of the image error

$$\arg\min_{N,\rho} \|I - \mathcal{M}(N, \rho)\| \tag{5.1}$$

where I is an input image. However, this optimization is not straightforward due to the large amount of parameters and their non-linear relationship. For a related energy based on point light sources, Goldman *et al*. [Goldman05] solve for materials and normals in an iterative procedure. We adopt a similar approach, but introduce the relative light intensities as optimization parameters and also account for the contribution of the sky.

5.1 Problem Statement and Overview

Given a large set of input images $I_1, \ldots, I_M$ of an outdoor scene observed by a static webcam, our goal is to recover the surface orientation and reflectance properties of objects in the scene. More specifically, we represent reflectance as a weighted sum of

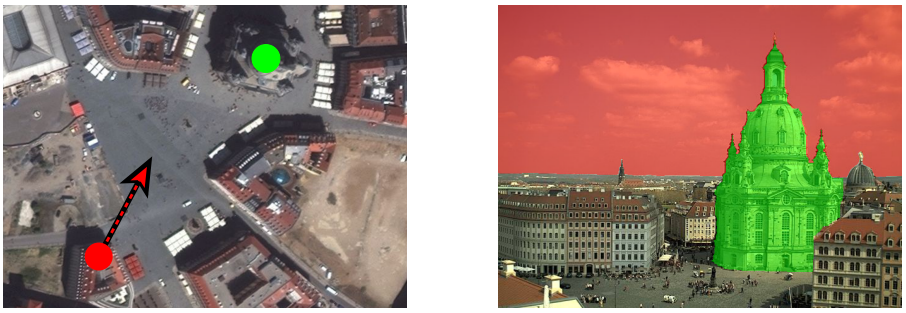


Figure 5.1: The geographic context of the scene with camera (red marker) and target object (green marker) on the left (source: Google Maps), and an example image from the camera with sky mask (red region) and object mask (green region) on the right.

basis materials per pixel, see [Lensch03]. Thus, we have to estimate the basis materials $f_1, \ldots, f_K$ as well as weight maps $\gamma_1, \ldots, \gamma_K : \{1, \ldots, W\} \times \{1, \ldots, H\} \to [0, 1]$ that describe the per-pixel mixture of materials. The number of basis materials K is the only information about reflectance that has to be supplied by the user. It is worth noting that the basis materials in our formulation are not "real" BRDFs—even if we use the term interchangeably—in the radiometric sense. They rather represent the most plausible explanation of scene reflectance. We model the camera as orthographic with a resolution of $W \times H$ and assume that its GPS coordinates are at least approximately known. Furthermore, we require the images to be equipped with a time-stamp.

We also need the camera orientation in a geo-referenced coordinate system. This is determined automatically during calibration by exploiting a sky model which implies that a portion of the sky has to be visible in the images. We facilitate the calibration by providing a mask that labels the sky region. To guide the image selection, we specify a second mask that segments the object of interest.

Our approach consists of two separate and independent steps (see Figure 5.2). We first select two subsets of "good" images according to the criteria discussed in Section 5.3. We then propose a calibration pipeline that removes pixel misalignment caused by camera movement, estimates the afore-mentioned camera orientation, and computes the position of the Sun for each selected image. Additionally, the camera response function is recovered and the pixel intensities are linearized. The second set of images together with per-image sun positions are the input to the second stage: an iterative optimization of the normal map $N : \{1, \ldots, W\} \times \{1, \ldots, H\} \to \mathcal{S}$, the basis materials f_k, and the mixture maps γ_k. This step exploits our image creation model for outdoor scenes which also introduces relative light intensities to account for exposure variation between images.

5.2 Image Creation Model

We will first describe the image formation model that underlies our approach because it motivates several of the design choices. The definitions in this section are relative to the camera coordinate system.

The luminance leaving a surface point x in the direction $D_{x,p}$ towards a point p

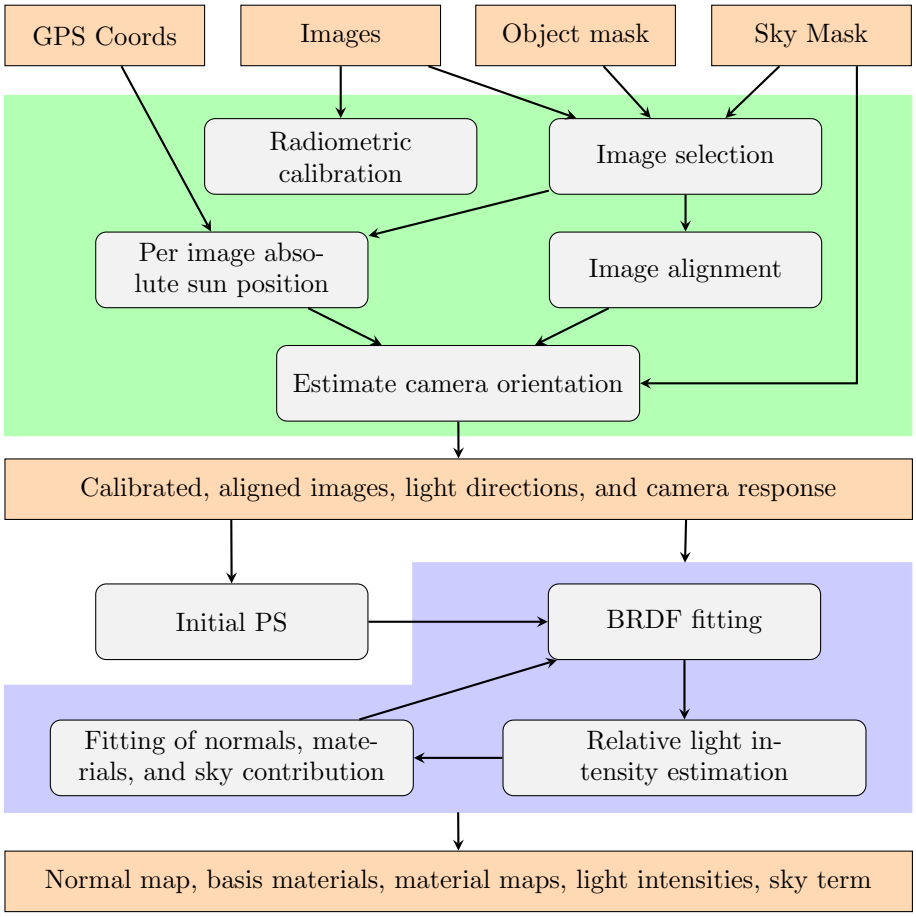


Figure 5.2: Algorithm overview. We first select suitable images and calibrate the camera (green) before we recover the final normal map, basis materials, corresponding material maps, and light intensities (purple).

on the image plane is given by Equation (2.16) as

$$\tilde{L}(D_{x,p}) = \int \rho_x(D_{x,p}, D_{in}) \cdot \langle n,\ D_{in} \rangle \cdot L_{s,x}(D_{in})\, d\omega_{D_{in}} \tag{5.2}$$

where $L_{s,x}$ is the distribution of incoming radiance, $d\omega_{D_{in}}$ is the infinitesimal surface element on a sphere, and ρ_x is the BRDF. Assuming there is no participating medium in the optical path and ignoring the details of interaction with the lens, the *incoming* luminance at p from the direction of x is equal to $\tilde{L}$. It gets transformed into pixel intensities I at discrete positions according to the camera model defined in Section 2.3.2. For now, we assume the response curve f to be linear. I is thus proportional to the luminance $I \propto \tilde{L}$. Dropping the spatial indices, we define $\rho(n, D) := \rho_x(D_{x,p}, D_{in}) \cdot \langle n,\ D_{in} \rangle$ and obtain the general formula

$$I = \beta \cdot \int \rho(n, D) \cdot L_{s,x}(D)\, d\omega_D \tag{5.3}$$

where β is the factor of proportionality which encompasses, for example, the exposure time and aperture size.

Different representations of reflectance are common in computer graphics. Some assume a parametric form of the BRDF, *e.g.* [Torrance67, Phong75, Blinn77, Ward92], others use measured data, *e.g.* [Matusik03], or linearize the BRDF through a decomposition into (wavelet) basis functions, *e.g.* [Haber09]. We model the function ρ for a scene point observed at pixel p in color channel c as a linear combination $\rho = \sum_{k=1}^{K} \gamma_{p,k} f_{k,c}$ of *basis materials* $f_{k,c}$ similar to Lensch *et al.* [Lensch03]. The materials $f_{k,c}$ are represented by a parametric model that Ward [Ward92] developed based on empirical measurements. In our case, it takes the form

$$f_{k,c}(n, D) = \left(\frac{\alpha_{\text{diff},c}}{\pi} + \frac{\alpha_{\text{spec},c}}{4\pi\,\tilde{\alpha}^2 \sqrt{\langle n,\ D\rangle\, n_z}} \cdot e^{-\left(\frac{\tan\langle n,\ w\rangle}{\tilde{\alpha}}\right)^2} \right) \cdot \langle n,\ D\rangle \tag{5.4}$$

where $w = \frac{D+(0,0,1)}{\|D+(0,0,1)\|}$ is the halfway vector between the light and viewing directions. The definition is invariant with respect to a rotation of the coordinate system because it depends only on scalar products of the input vector. The parameter $\tilde{\alpha}$ represents the surface roughness and controls the breadth of the highlight. We use individual coefficients $\alpha_{\text{diff},c}, \alpha_{\text{spec},c}$ for the diffuse and specular contribution and a single roughness parameter for all color channels. The advantage of this model is that it has only seven parameters $\alpha = (\alpha_{\text{diff},RGB}, \alpha_{\text{spec},RGB}, \tilde{\alpha})$ which also have an intuitive interpretation.

The second part of Equation (5.3) that we have to discuss is the incoming luminance distribution $L_{s,x}$. We assume that $L_{s,x}(D) = L_s(D) \cdot V(x, D)$ can be factored into an overall distribution L_s that is equal for all surface points—which represents a distant light assumption—and a visibility function V that encodes whether luminance from direction D towards x is blocked. Since each x corresponds to a pixel p, we may also write $V(x, D) = V_p(D)$.

So far, the model is pretty general, and similar ones have been used in many computer graphics applications, *e.g.* [Yu99, Goldman05]. However, the degrees of freedom necessary for arbitrary illumination make the inverse rendering problem intractable. We solve this issue by exploiting prior knowledge in the form of an outdoor lighting model. Illumination from the Sun and sky is studied in many fields such as energy engineering, architecture, environmental sciences, and computer graphics. Accordingly, sky models with different requirements on generality and accuracy have been developed [Brunger93, Perez93, Preetham99, Darula02]. For our purposes, we propose a simple model that assumes a clear sky with a bright Sun as the primary light source. We separate the luminance into a sun and sky contribution $L_s = L_{\text{Sun}} + L_{\text{Sky}}$. The linearity of Equation (5.3) yields the intensity at a pixel p and color channel c in the i-th image:

$$I_{i,p,c} = I_{\text{Sun},i,p,c} + I_{\text{Sky},i,p,c}. \tag{5.5}$$

We model the Sun as a point light source $\delta(D - D_{i,\text{Sun}})$ from direction $D_{i,\text{Sun}}$ with intensity $\tilde{l}_{i,c}$. The delta peak then collapses the integral from Equation (5.3) into the simpler form

$$I_{\text{Sun},i,p,c} = \beta_i \cdot \tilde{l}_{i,c} \sum_{k=1}^{K} \gamma_{p,k} f_{k,c}(n_p, D_{i,\text{Sun}}) V_p(D_{i,\text{Sun}}) \tag{5.6}$$

with the visibility function corresponding to cast shadows. Furthermore, we approximate the sky as a spatially uniform light source of intensity $\widetilde{S}_{i,c}$. Now, V_p encodes the

portion of the sky that is visible—similar to ambient occlusion, see [Pharr12, Chapter 17]:

$$I_{\text{Sky},i,p,c} = \beta_i \cdot \widetilde{S}_{i,c} \int \sum_{k=1}^{K} \gamma_{p,k} f_{k,c}(n_p, D) V_p(D)\, d\omega_D \tag{5.7}$$

$$= \beta_i \cdot \widetilde{S}_{i,c} \cdot \text{const}_p\,. \tag{5.8}$$

Lastly, if we assume that the intensity of the sky scales linearly with the sun intensity, *i.e.* $\widetilde{S}_{i,c} \propto \tilde{l}_{i,c}$ independent of i, we can factor this into per-image and per-pixel terms

$$I_{\text{Sky},i,p,c} = \left(\beta_i \cdot \tilde{l}_{i,c}\right) \cdot S_{p,c}. \tag{5.9}$$

For webcams, camera parameters such as exposure time and gain are usually unknown and may vary over time. This means the observed intensity in the image might differ from the true scene luminance by the unknown scalar factor β_i. We incorporate these effects by considering relative light intensities

$$l_{i,c} := \frac{\beta_i \cdot \tilde{l}_{i,c}}{\beta_1 \cdot \tilde{l}_{1,c}} \tag{5.10}$$

normalized to the first image. Replacing the absolute intensities $\tilde{l}_{i,c}$ in Equation (5.6) and Equation (5.9) gives the final image creation model

$$I_{i,p,c} = l_{i,c} \left(\sum_{k=1}^{K} \gamma_{p,k} f_{k,c}(n_p, D_{i,\text{Sun}}) V_p(D_{i,\text{Sun}}) + S_{p,c} \right). \tag{5.11}$$

5.3 Image Selection

Our image creation model can only be an approximation to reality, and thus some images will violate its assumptions. Including these outliers in the reconstruction will disturb the optimization and produce unsatisfying results. Moreover, there will be a lot of redundancy even in the images that fit the model, *e.g.* if taken only a few minutes apart. Including all those in the reconstruction will use lots of computational resources without increasing the information content. We would therefore like to select a small subset of 30 to 50 "suitable" images out of the 20 000 and more that we collect.

Since manual selection is a tedious task, we present a novel algorithm to determine this subset automatically. First, we need to define and formalize what "suitable" means in this context. Ideally, we would like to use images where the object is well exposed and directly illuminated by the Sun on a cloudless sky. Thus, we have to look at both, the appearance of the object and the appearance of the sky. Overcast conditions are not covered by our model and should therefore be excluded. Figure 5.3 gives some examples of object and sky appearance that occur in webcam images.

We found a plausible good weather indicator in the "blueness" of the sky. To this end, we compute the average color $\bar{S}_{R,G,B}$ in the sky mask for each of the RGB color channels and define

$$B_{\text{sky}} = \bar{S}_B - \max(\bar{S}_R, \bar{S}_G) \tag{5.12}$$

Figure 5.3: Exemplary webcam images for the *church* dataset [Webb] demonstrating the large variation in object and sky appearance.

as a measure of blueness. This indicator helps to exclude images with a completely overcast sky, but it neglects the illumination of the actual object. Also, it does not consider the presence of small clouds or haze. If these are not in front of the sun, they will not influence I_{Sun} and only slightly impact the assumed uniform intensity of the sky. We view such images as suitable for reconstruction. But if a cloud actually occludes the sun, it scatters the light and turns the point light source into an area light not covered by our model. The consequence is a lower contrast of the object compared to images under direct illumination. This can even occur if B_{sky} is high because the Sun and cloud might be behind the camera. We measure the variance V_{obj} of all intensity values in the object mask $\mathcal{M}_{\text{obj}}$ and the average gradient magnitude

$$G_{\text{obj}} = \frac{1}{|\mathcal{M}_{\text{obj}}|} \sum_{p \in \mathcal{M}_{\text{obj}}} \|\nabla I(p)\| \tag{5.13}$$

as indicators of high contrast, well exposed images.

It is not sufficient to look only at the suitability of individual images. For photometric stereo to be well-defined, light directions need to exhibit sufficient variation. If all directions actually lie on a plane, such as the sun positions throughout a single day as considered by Sunkavalli *et al.* [Sunkavalli08], only the projection of the normal onto that plane can be recovered. We therefore capture images over the course of at least 6 months and define a penalty function on the time stamps of image pairs:

$$P^{i,j} := e^{-\frac{(x_j - x_i)^2}{2\sigma_x^2} - \frac{(y_j - y_i)^2}{2\sigma_y^2}}. \tag{5.14}$$

Here, x_j and x_i are the day of the year for image j and i, and y_j and y_i are the minute of the day for image j and i respectively. The 2D Gaussian discourages pairs taken at the same time of day—corresponding to similar azimuth angles—or taken at consecutive days—corresponding to similar zenith angles. Figure 5.4 demonstrates that effect. We achieved good results with $\sigma_x = 10$ and $\sigma_y = 30$, which corresponds to a Gaussian spread of 10 days over the year and 30 minutes over the day.

Before we run the actual selection process, we exclude images that are obviously outliers. They are rejected immediately if more than 10 % of the pixels in either the

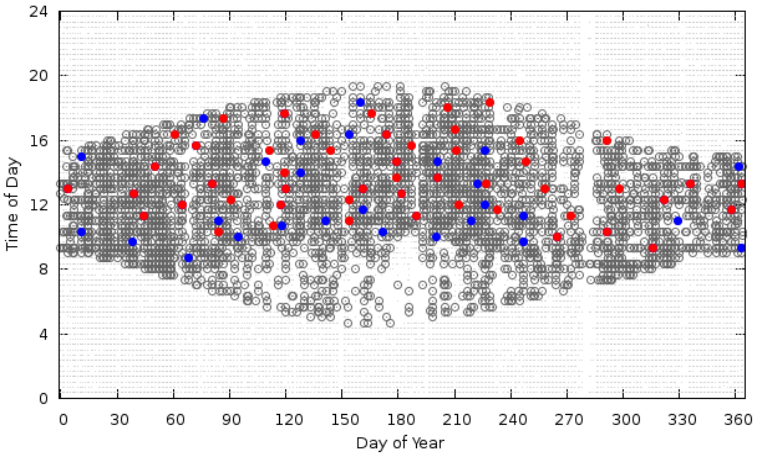


Figure 5.4: Penalizing similar sun positions on the *church* dataset. The plot shows selected clear sky images (blue) and selected images for photometric stereo (red), all candidate images (dark gray circles) and all available non-overexposed images (light gray dots).

whole image or the object region are overexposed. To discard night-time images and extreme lighting conditions during sunrise or sunset, we remove images where the Sun is close to the horizon. More precisely, we disregard images where the sun zenith angle is larger than 85°. We describe how to obtain this information from the time stamp in Section 5.4.3. We also exclude underexposed images. To this end, we compute the median intensity of the sky region I_{sky} and the 75th percentile intensity of the object region I_{obj} as defined by the masks. The latter rewards images where the object is well illuminated but also allows the object to be partially shadowed. We then discard 50 % of the images with lowest score $S_I = I_{\text{sky}} + I_{\text{obj}}$.

After the initial pruning, we perform the actual image selection. We combine the weak indicators introduced before into a score that takes individual images into account as well as the temporal distribution within the selected set:

$$S^i_{\text{PS}} = G^i_{\text{obj}} \cdot B^i_{\text{sky}} \cdot V^i_{\text{obj}} \cdot P^i. \tag{5.15}$$

Each of the quantities B^i_{sky}, G^i_{obj}, and V^i_{obj} is normalized with respect to its minimum and maximum value over all images. P^i is initially set to one for all images. In a greedy search, we then select the image with best score S^i_{PS} and include it in the final set. Afterward, we update P^j for each not yet selected image according to

$$P^j = P^j - P^j \cdot P^{i,j} \tag{5.16}$$

which penalizes similar sun positions. We iterate this selection and update process until the required number of images is selected.

The described selection process is geared towards images for photometric reconstruction. We need another subset of images for the calibration of absolute camera orientation explained in Section 5.4.4. This exploits a complex sky model and requires an adjusted definition of a “suitable” image. The calibration is not concerned with object contrast but needs the visible sky region to be cloudless. An example of these differences is shown in Figure 5.5. The luminance in such a sky varies smoothly, and

Figure 5.5: Image selection for the *church* dataset [Webb]. *Left*: Two example images with clear sky selected for calibration. *Right*: Two of the photometric stereo images that focus more on the object contrast than on a clear sky.

clouds then introduce strong gradients. Accordingly, low values of the average gradient magnitude in the sky mask G_{sky} indicate suitable images. Thus, we replace the object gradient in Equation (5.15) to obtain the score

$$S^i_{\text{clearsky}} = (1 - G^i_{\text{sky}}) \cdot B^i_{\text{sky}} \cdot V^i_{\text{obj}} \cdot P^i. \tag{5.17}$$

The same iterative construction of the image set is applied as for S^i_{PS}.

5.4 Webcam Calibration

5.4.1 Image Alignment

Outdoor conditions such as strong winds can cause the webcam to shake, resulting in small camera motions. We apply a subpixel alignment step since even subtle misalignment has serious impact on the reconstruction quality. The dramatic variations in image appearance disqualifies most naive methods. We use gradient images to achieve a certain robustness to appearance changes. However, aligning such images directly using Lucas-Kanade [Lucas81] fails since the gradients are heavily influenced by varying lighting and shadows. We instead align the gradient images to the average gradient image, calculated from all input images, similar to Jacobs *et al.* [Jacobs09]. Figure 5.6 shows the alignment in pixel coordinates for about 50 images from the *church* dataset.

5.4.2 Radiometric Calibration

As discussed in Section 2.3.2, it is important for photometric methods to operate on linear intensity values. Obtaining those requires a radiometric calibration which is difficult to perform on an uncontrolled camera. We make use of the approach

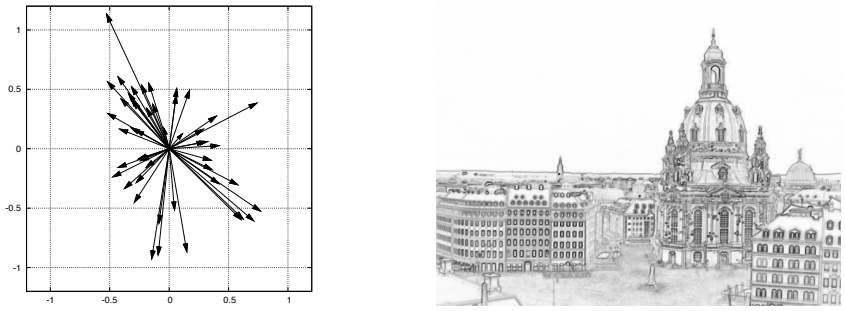


Figure 5.6: Camera shake. *Left*: Each arrow corresponds to an image that has been aligned along the direction of its arrow. The axes show the distance of the alignment in pixels. One can see that alignment is very subtle, barely more than a single pixel. *Right*: The average gradient image used for alignment.

developed by Kim *et al.* [Kim08a] that is specialized for outdoor scenes. First, we briefly summarize their idea and then relate our image formation model to theirs.

Kim *et al.* model the reflected luminance as the product of an exposure value β_i, a diffuse albedo α_p, and the illumination term $M_{p,i}$ which includes directional and ambient lighting:

$$L_{i,p} = \beta_i \alpha_p M_{p,i}. \tag{5.18}$$

They find that the relationship between two surfaces points, which have the same normal and are either both in shadow or both lit, depends solely on the albedo ratio

$$\frac{L_{i,p}}{L_{i,q}} = \frac{\beta_i \alpha_p M_{p,i}}{\beta_i \alpha_q M_{q,i}} = \frac{\alpha_p}{\alpha_q}. \tag{5.19}$$

To detect such points, they perform a clustering of pixels based on similar appearance over multiple images. The clusters are then filtered to prune outliers. We use the implementation provided by the authors and refer to the respective article [Kim08a] for further details on the preprocessing.

The linear luminance L gets transformed by the camera's response function f into pixel values $I = f(L)$. A common approach to estimate response functions from a small set of noisy samples is to model the problem as superposition of basis functions and then estimate the coefficients of the linear system which arises. Kim *et al.* use the popular linearization of the inverse logarithmic response $G := \log \circ f^{-1}$ introduced by Grossberg and Nayar [Grossberg03, Grossberg04]:

$$G(I) = G_0(L) + \sum_j \tau_j h_j(L) \tag{5.20}$$

where the basis functions G_0, h_j have been obtained through principal component analysis of a large database of known response curves. For pixels within the same cluster, the difference

$$G(I_p) - G(I_q) = G_0(I_p) - G_0(I_q) + \sum_j \tau_j \left(h_j(I_p) - h_j(I_q)\right) \tag{5.21}$$

is equal to

$$G(I_p) - G(I_q) = \log \frac{f^{-1}(I_p)}{f^{-1}(I_q)} = \log \frac{L_p}{L_q} = \log \frac{\alpha_p}{\alpha_q} = \log \alpha_p - \log \alpha_q =: a_{p,q} \tag{5.22}$$

obtained from Equation (5.19). In combination, this yields a linear system in the coefficients $a_{p,1}, \ldots, a_{p,Q}, \tau_1, \ldots, \tau_J$ for all Q pixels in the cluster according to

$$a_{p,q} - \sum_j \left(h_j(I_{i,p}) - h_j(I_{i,q})\right) \cdot \tau_j = G_0(I_{i,p}) - G_0(I_{i,q}). \tag{5.23}$$

The resulting equations for different clusters and different images I_i can be stacked into a single large system. The solution then yields the optimal $\tau_1, \ldots, \tau_J$ which define the inverse response function G. Exemplary response curves recovered for two of our datasets are shown in Figure 5.7.

Assuming purely diffuse basis materials, our image creation model can be interpreted as a special case of Equation (5.19) and does not violate the assumptions made by Kim *et al*. Our BRDFs then attain the form

$$f_k(n, D) = \frac{\alpha_{\text{diff},k}}{\pi} \cdot \langle n,\ D \rangle \tag{5.24}$$

and only differ in albedo. Inserting into Equation (5.6) and Equation (5.8) yields an incoming luminance

$$L_{i,p} = \beta_i \left(\tilde{l}_i A_p \langle n,\ D_{i,Sun} \rangle V_p(D_{i,Sun}) + \tilde{S}_i A_p \int \langle n,\ D \rangle V_p(D)\, d\omega_D \right) \tag{5.25}$$

where $A_p = \frac{1}{\pi} \sum_{k=1}^{K} \gamma_{p,k} \cdot \alpha_{\text{diff},k}$. For two surface points p and q that have the same normal and the same visibility function V, the ratio

$$\frac{L_{i,p}}{L_{i,q}} = \frac{\beta_i A_p \left(\tilde{l}_i \langle n,\ D_{i,Sun} \rangle V(D_{i,Sun}) + \tilde{S}_i \int \langle n,\ D \rangle V(D)\, d\omega_D \right)}{\beta_i A_q \left(\tilde{l}_i \langle n,\ D_{i,Sun} \rangle V(D_{i,Sun}) + \tilde{S}_i \int \langle n,\ D \rangle V(D)\, d\omega_D \right)} = \frac{A_p}{A_q}. \tag{5.26}$$

then no longer depends on i and is constant for all images. Thus, replacing α_p in Equation (5.19) with the weighted sum of diffuse basis materials A_p yields the same starting point as required by Kim *et al*.

5.4.3 Sun Position

We store the time of download along with each image which makes us independent of meta-information provided by the camera that might be wrong or simply missing. Then, the sun direction $D_{i,\text{Sun}}$ can be computed from the time-stamp. The Earth moves on an ellipse around the Sun and also rotates about its own axis—which is slightly tilted with respect to the orbital plane. Thus, the position of the Sun relative to the Earth can be computed at any point in time. If we know the position of the camera in a geo-referenced coordinate system, we can transform the position of the Sun from the geocentric into a local coordinate system with one axis pointing north and one pointing upwards. The details of these computations and the astronomical background are beyond the scope of this thesis. We apply the implementation of a Solar Position

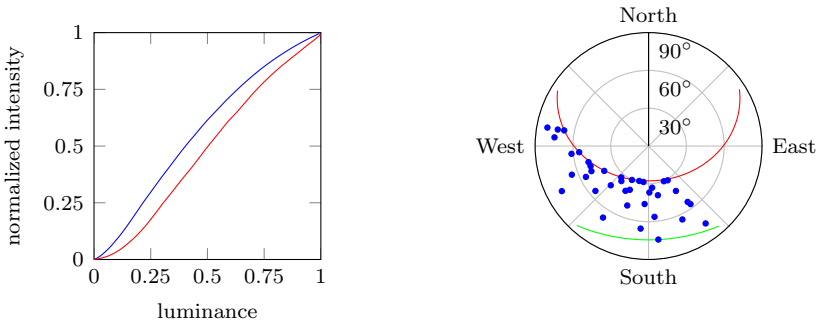


Figure 5.7: Webcam calibration. *Left:* Recovered response curve for the *church* (blue) and *tower* (red) dataset. *Right:* Sun positions for the *church* dataset. Directions towards the east are missing because the automatic selection shown in Figure 5.4 only chose images after 8am.

Algorithm provided by the National Renewable Energy Laboratory [Reda04] to recover the zenith and azimuth angles $(\tilde{\theta}, \tilde{\phi})$ in the local coordinate system. Figure 5.7 (right) shows the sun positions recovered for the *church* dataset. The red and green lines indicate the solstices and define the maximal spread of directions possible at that location.

5.4.4 Camera Pose

The model in Section 5.2 is defined in terms of camera coordinates. We have to find a transformation from the local coordinate system into the camera coordinate system to obtain the appropriate direction $D_{i,\text{Sun}}$. This is equivalent to determining the camera's viewing direction $\theta_{\text{cam}}, \phi_{\text{cam}}$ in the geo-referenced, local coordinate system. A solution to this problem has been proposed by Lalonde *et al.* [Lalonde10], and we will briefly summarize their approach.

They exploit the sky model by Perez *et al.* [Perez93] which, given the luminance $S(0,0)$ at the zenith, predicts the luminance from direction (θ, ϕ) as

$$S(\theta,\phi) = S(0,0)\,\frac{(1+\sigma_1\, e^{\sigma_2/\cos\theta})\cdot(1+\sigma_3\, e^{\sigma_4\chi}+\sigma_5\cos^2\chi)}{(1+\sigma_1\, e^{\sigma_2})\cdot(1+\sigma_3\, e^{\sigma_4\tilde{\phi}}+\sigma_5\cos^2\tilde{\phi})} \tag{5.27}$$

where $\chi(\theta,\phi,\tilde{\theta},\tilde{\phi})$ is the angle between the desired direction and the direction of the Sun. The parameters $\sigma_{1,\ldots,5}$ control the sky appearance and correspond to overcast, uniform, clear, *etc.*, conditions. Lalonde *et al.* relate the angles (θ, ϕ) in the local coordinate system to pixel coordinates $p = (u, v)$ in the camera coordinate system. Inserting their transformation into Equation (5.27) yields a function

$$S(0,0)\cdot W(\theta_{\text{cam}}, \phi_{\text{cam}}, F, u, v, \tilde{\theta}, \tilde{\phi}) \tag{5.28}$$

to predict the luminance at pixel p. The unknown focal length F, viewing direction $(\theta_{\text{cam}}, \phi_{\text{cam}})$, and zenith luminance $S(0,0)$ can then be estimated by comparison with the observed image intensities

$$\arg\min \sum_i \sum_{u,v} \Big(I_{i,p} - S_i(0,0)\cdot W(\theta_{\text{cam}}, \phi_{\text{cam}}, F, u, v, \tilde{\theta}_i, \tilde{\phi}_i) \Big)^2 . \tag{5.29}$$

We refer to the corresponding publication [Lalonde10] for further details and a pointer to the publicly available source code.

5.4.5 Shadow Detection

Shadows can provide a useful cue to infer scene structure [Daum98]. In photometric stereo, they are, however, typically a source of additional errors. This is especially true for outdoor scenes as cast shadows dramatically change the appearance of the images.

To detect a shadowed pixel, we use the method originally proposed by Sunkavalli *et al.* [Sunkavalli07]. The basic idea is that the pixel intensity $I_{i,p}$ of pixel p in the i-th image will be significantly lower when in shadow than when exposed to direct sunlight. If we do not observe an intensity difference of at least a factor of 1.4 between the highest and the lowest pixel value, we assume that the pixel was shadowed in all images. For each other pixel p, we first calculate the median value $m_{\min}$ of the n smallest intensity values. The pixel is detected as shadowed in image i if its intensity $I_{i,p}$ in that image is smaller than $\tau \cdot m_{\min}$. We use $n = 10\%$ of the total number of images and $\tau = 1.5$ in our examples.

5.5 Reconstruction

Our algorithm optimizes for the relative light intensities $l_{i,c}$ in each image, the basis materials $f_{1,c}, \ldots, f_{K,c}$ in the scene, the surface orientation n_p, sky light contribution $S_{p,c}$, and material mixing coefficients $\gamma_{p,k} \geq 0$ at each pixel. Sun visibility $V_p(D_{i,\text{Sun}})$ is handled by the shadow detection and we replace it by introducing the sets $\mathcal{I}_p$ of images such that p is not in shadow. Given P pixels within the mask and M images, we obtain

$$\big(2 + 3 + K\big) \cdot P + 3 \cdot \big(M - 1\big) + 7 \cdot K \tag{5.30}$$

unknown variables in total—ignoring shadows. From these variables, we can render a synthetic image according to Equation (5.11) and compare its intensities against the observed images $\hat{I}$. Employing the L_2 error leads to the minimization problem:

$$\arg\min \sum_{p=1}^{P} \sum_{i \in \mathcal{I}_p} \sum_{c=1}^{3} \big(\hat{I}_{i,p,c} - I_{i,p,c}\big)^2. \tag{5.31}$$

Each pixel in each image introduces one constraint per color channel. The total number of equations is thus $3 \cdot P \cdot M$. We need at least

$$M > \frac{(2 + 3 + K)\, P - 3 + 7\, K}{3\,(P - 1)} \tag{5.32}$$

images to have sufficient constraints for an overdetermined system.

5.5.1 Initialization

The above optimization problem is non-linear and has thousands of unknowns. It is important to choose a starting point that is already close to the final solution. During

Figure 5.8: Automatic initialization. *Left to right*: An input image of the *tower* dataset [Webd], detected shadow regions to determine $\mathcal{I}_p$, recovered object albedo from Lambertian photometric stereo, and selected points to initialize the intensity estimation.

the initialization, we ignore the sky term $S_{p,c}$ and set it to zero. In a first step, we assume constant light intensities and treat the whole scene as Lambertian, ignoring the errors that will arise if a more complex material is present. In our experiments, we found that all scenes contained enough points that conform to this simplification and result in decent initial estimates of surface normals and albedo. With this information, we find an initialization of the relative light intensities and use them to re-estimate the normals, which considerably improves the first initialization. Finally, we cluster the resulting albedos and fit "pure" basis materials to get an initialization of the per-pixel reflectance properties. We now describe each of these steps in more detail and visualize intermediate results in Figure 5.8.

Classical Photometric Stereo: We solve for the normal $n_{p,c}$ and albedo $\rho_{p,c}$ in each color channel by minimizing

$$E(\rho_{p,c}, n_{p,c}) = \frac{1}{|\mathcal{I}_p|} \sum_{i \in \mathcal{I}_p} \left(I_{i,p,c} - \rho_{p,c}\, l_{i,c}\, \langle n_{p,c},\ D_i \rangle\right)^2 . \tag{5.33}$$

The relative intensities $l_{i,c}$ are set to 1.0 during initialization. We ignore the pixel if $\mathcal{I}_p$ contains too little variation in the light directions, indicated by $\min_{i,j} \langle D_i,\ D_j \rangle > \cos(37°)$, or if the least squares error is too high. For valid pixels, we select the normal from the color channel with the lowest error.

Relative Light Intensities: For a first estimate of the relative light intensities, we employ the approach by Hayakawa [Hayakawa94]. To overcome certain ambiguities, this method needs at least six surface points with similar albedos and sufficient variation in normal direction. We determine these points automatically by clustering all pixels into four albedo clusters according to the albedo estimate from Equation (5.33). For the most frequent albedo, we then cluster the normals of the set into 30 normal

clusters. Finally, we select eight normals from different clusters and ensure that the corresponding pixels are almost never in shadow (see Figure 5.8 for an example). In both steps, the clustering is done using Expectation Maximization for a Gaussian mixture model.

Initial Material Estimation: Following Goldman *et al.* [Goldman05], we use the albedos to compute an initial distribution of the fundamental materials in the scene. The number K of fundamental materials is chosen beforehand. (Typically, two or three materials are sufficient for accurate material reconstruction.)

While Goldman *et al.* suggest to cluster the albedos in the HSV color space, we found that clustering in sRGB color space gives better results for our datasets. Again, the clustering is based on a mixture of Gaussians. For each pixel, we assign its cluster weights as initial material combination $\gamma_{p,1}, \ldots, \gamma_{p,K}$ (normalized to $\gamma_{p,1} + \ldots + \gamma_{p,K} = 1$). Given these initial estimates for the mixing coefficients, we now find good initializations for the parameters of each material. For each $k \in \{1, \ldots, K\}$, we build a set of pixels $\mathcal{K}_k$ that represent the *pure* material. We do this by selecting pixels with the mixing coefficient for that material at least ten times greater than the others. Based on these pure pixels, we fit the BRDF parameters using non-linear Levenberg-Marquardt optimization [Press92], minimizing:

$$E(\alpha_k) = \sum_{c,p \in \mathcal{K}_k, i \in \mathcal{I}_p} \left(I_{i,p,c} - l_{i,c} f_c(n_p, D_{i,\mathrm{Sun}}, \alpha_k)\right)^2 . \tag{5.34}$$

5.5.2 Iterative Refinement

We found that an optimization based directly on Equation (5.31) often gets stuck in local optima. Instead, we split it into three steps which we iterate to refine the material parameters, the light intensities, and the per-pixel estimates (normal map, sky contribution, and mixing maps). An overview of the entire algorithm is shown in Figure 5.2. As each subproblem decreases the objective function, the algorithm is guaranteed to converge but might result in a local optimum. Depending on the scene, about 50 iterations were sufficient to reach convergence. Figure 5.9 demonstrates this for the *church* dataset. After each of the following steps, we update the current intensity estimate of pixel p in image i and color channel c, which we denote

$$e_{i,p,c} = l_{i,c} \left(\textstyle\sum_{k=1}^{K} \gamma_{p,k} f_c(n_p, D_i, \alpha_k) + S_{p,c} \right) . \tag{5.35}$$

This estimate corresponds to rendering an image with the current parameter set as shown in Figure 5.9.

Material Fitting: In Equation (5.34), we optimized parameters for each material separately. This already provides us with a good initial estimation. Now, we find the optimal parameters for all materials simultaneously, i.e., for the concatenation α of all parameter vectors α_k and not restricted to *pure* pixels $\mathcal{K}_k$. Given the current estimate for the normal map and per-pixel material weights, we minimize

$$E(\alpha) = \textstyle\sum_{i \in \mathcal{I}_p, p, c} \left(I_{i,p,c} - e_{i,p,c}\right)^2 . \tag{5.36}$$

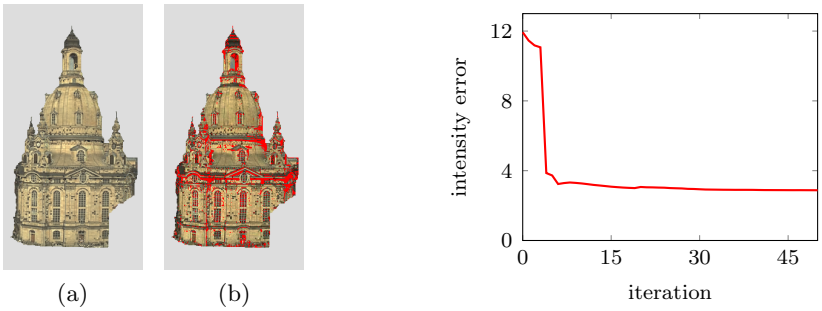


Figure 5.9: *Left:* One of the renderings (a) that arises during optimization. This estimate is compared to the true image (b, with mask and shadow detection applied). *Right:* The error decreases quickly. We stop the optimization after 50 to 150 iterations depending on the dataset.

Light Intensity Optimization: To improve the relative intensities during our optimization, we analytically solve for the best intensity update $U_{i,c} = l_{i,c}^{\text{new}} / l_{i,c}^{\text{current}}$ in every image, while keeping all other variables fixed. We want to minimize

$$E(U_{i,c}) = \sum_p \left(I_{i,p,c} - U_{i,c} e_{i,p,c}\right)^2 . \tag{5.37}$$

Setting the derivative to zero yields the intensity update

$$0 = \frac{\partial E}{\partial U_{i,c}} = -2 \sum_p e_{i,p,c} \left(I_{i,p,c} - U_{i,c} e_{i,p,c}\right) \quad \Longleftrightarrow \quad U_{i,c} = \frac{\sum_p e_{i,p,c} I_{i,p,c}}{\sum_p e_{i,p,c}^2}. \tag{5.38}$$

Applying the update may lead to a reference intensity $l_{1,c}^{\text{new}} = U_{1,c} \cdot l_{1,c}^{\text{current}} \neq 1$. We re-normalize the intensities before proceeding.

Material and Normal Map Optimization: The next step calculates the material weight maps, the per-pixel sky contribution $S_{p,c}$, and the normal for each pixel while material parameters and light intensities are fixed. For each pixel, we minimize

$$\begin{aligned} &E(n_p, S_p, \gamma_{p,1}, ..., \gamma_{p,K}) \\ &\quad = \sum_c \left(\sum_{i \in \mathcal{I}_p} (I_{i,p,c} - e_{i,p,c})^2 + \lambda S_{p,c}^2 \right) \end{aligned} \tag{5.39}$$

with a Levenberg-Marquardt optimization. Note that the material weights are no longer restricted to sum to one. We also found that the optimization might run into local optima with wrong normals by compensating the error with implausibly high skylights $S_{p,c}$. We therefore include a penalty term to restrict its influence. In our experiments, we weight this term with $\lambda = 0.1$.

5.6 Evaluation

5.6.1 Synthetic Data

We first evaluate our algorithm on a synthetic dataset with known ground truth. The dataset shows a sphere consisting of two materials, diffuse gray and specular blue.

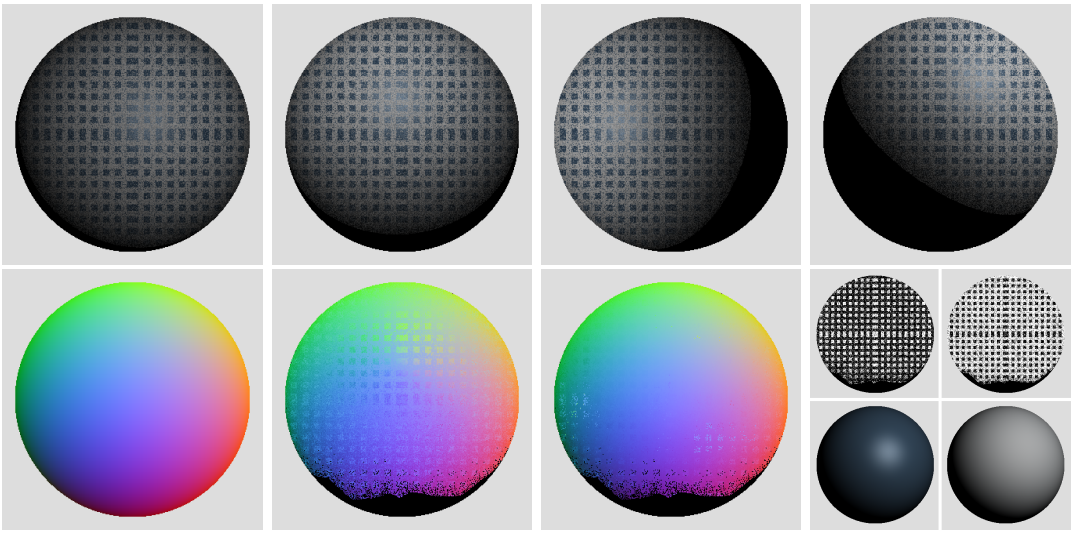


Figure 5.10: Synthetic input data. *Top*: Input images with varying light intensities and directions. The renderings contain two materials and randomly added sky contributions. *Bottom*: Ground truth normal map, initial normal map, normal map after optimization, and the final material maps with corresponding materials.

Additionally, each pixel has a 30 % chance to be a random mixture between both materials. We explicitly use an orthogonal projection and parallel light rays from a distant point light source to render ten images of the sphere. We also vary the light intensities in each image and add random indirect illumination per pixel corresponding to different local sky visibility functions. This setting reflects all aspects of our image formation model.

Figure 5.10 presents the result of our algorithm applied to these images. The optimized normals show a significant improvement over the purely diffuse photometric stereo used for initialization. Some small errors remain if the optimization runs into wrong local optima, but we are able to properly recover the material parameters and maps.

Figure 5.11 contains quantitative results. The angular deviation from the reconstructed normals to the ground truth is largest around the boundary where we observe the sphere under grazing angle. The mean angular error of all reconstructed normals is 2.76°. We chose the light positions to lie on the upper hemisphere to simulate solar illumination. Accordingly, errors increase in the lower part where only few intensity samples are available per pixel. Especially the sky contribution leads to ambiguous results in these regions because it dominates over the direct illumination. The plot in Figure 5.11 illustrates the performance of the light intensity estimation step. Since we recover only relative intensities, they should differ from the ground truth only by a global factor. This is verified by the fact that all samples lie on a line.

5.6.2 Webcam Data

We now demonstrate our technique on four outdoor datasets: the *tower* (≈20 k images), *church* (≈27 k images), *castle* (≈20 k images), and *sphere* (≈20 k images). For each of the datasets, we downloaded images every 20 minutes over the course of a

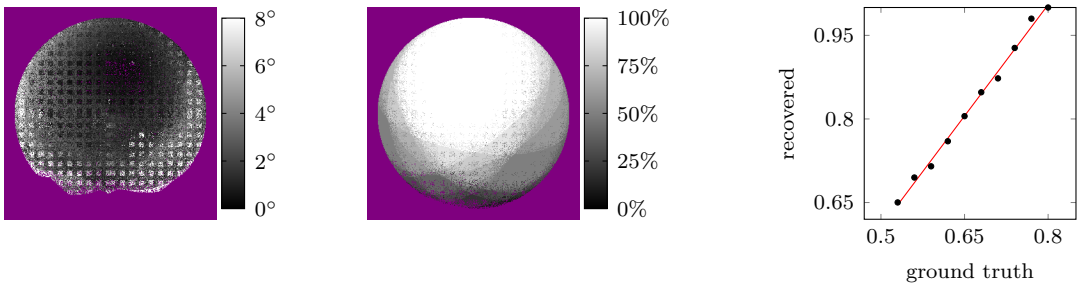


Figure 5.11: Quantitative results for synthetic data. *Left*: Angular deviation in degree from recovered normals to ground truth. Errors increase at grazing angles. *Middle*: Percentage of images for which the pixel is not in shadow. The light positions are chosen on the upper hemisphere to simulate solar illumination. *Right*: Estimated relative intensities are plotted against ground truth. The best fitting line (red) shows that results differ only by a global factor.

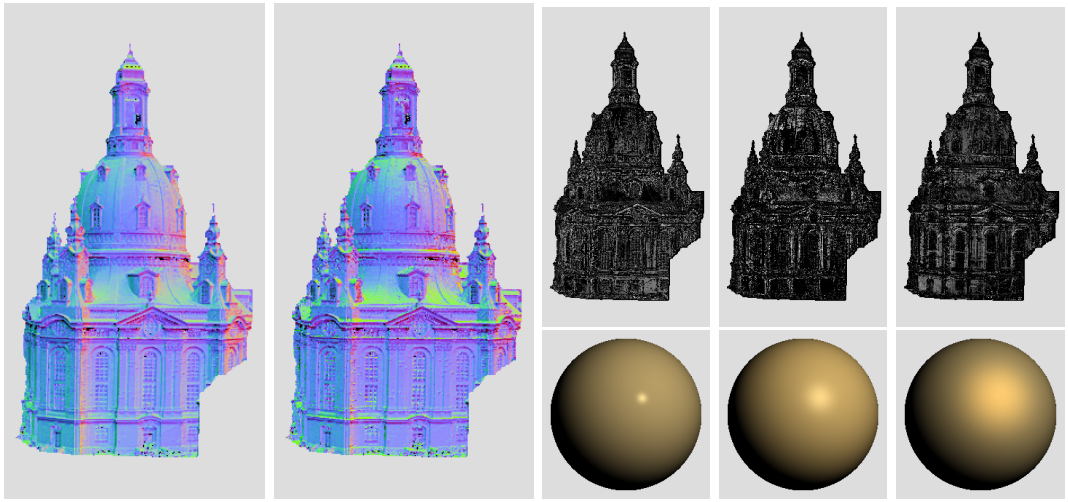


Figure 5.12: Results for the *church* dataset. *Left to right*: The initial normal map, the final normal map, and the three recovered BRDFs with corresponding material map.

year to get enough variation in the light directions. We then automatically selected 50 images for calibration and photometric stereo. An example of the selection for the *church* dataset is shown in Figure 5.5. We confirmed the calibration results visually by projecting the viewing direction to the ground plane as, *e.g.*, shown in Figure 5.1.

We first look at the results for the *church* dataset, which we have used repeatedly in this chapter. The initial normal map in Figure 5.12, reconstructed using the diffuse photometric stereo from Section 5.5.1, looks already promising, the final normal map exhibits more pronounced vertical directions of the normals in the dome region. The *church* is almost completely built from yellow sandstone, which is also reflected in the three recovered material maps. They indicate that almost every pixel is affected by all materials. Note that our method fits a fixed number of basis BRDFs that optimally explain the scene appearance. Since most scene points in this example have a similar albedo, the recovered materials essentially represent the specular variations.

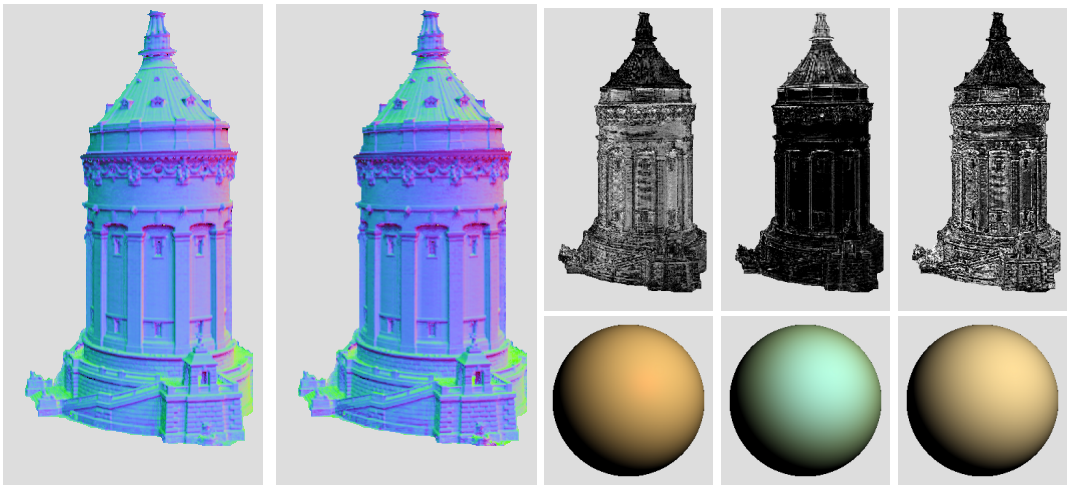


Figure 5.13: Results for the *tower* dataset. *Left to right*: The initial normal map, the final normal map, and the three recovered BRDFs with corresponding material map.

A more interesting example for the reflectance separation is given by the *tower*. It consists of a diffuse material (stone) and a more specular greenish material at the roof (corroded copper). Figure 5.13 shows that the different parts can be clearly distinguished in our reconstruction. Since we used three BRDFs for reconstruction, the appearance of the central part is again split into two similar materials. All pixels belonging to stone are made of a mixture of these brown and yellowish materials. Similar to the *church* dataset, the initial normal map is quite good and changes only slightly during optimization. This can be explained by the mostly diffuse behavior of both datasets.

It is difficult to evaluate the results quantitatively because ground truth data for outdoor scenes is hard to acquire. We exploit that the central part of the tower can be well approximated by a cylinder. Thus, to evaluate the performance of our technique, we render the normal map of a cylinder with corresponding radius and height as seen from a perspective camera. We guess the radius to be 11 m, the distance to the camera as 114 m, and determine its position by matching the rendering with one of the input images. This approach to generating reference data is not optimal but it allows us to get an impression of the quantitative performance of our technique. Figure 5.14 shows normals for one scanline from our reconstruction and the reference which we reproduce quite accurately. Most deviations occur at the far right and left where the surface is seen at grazing angles.

Especially in outdoor scenes, ambient light cannot be ignored entirely since it may lead to less pronounced normals. In Figure 5.14, we visualize the recovered sky term that shows the tower as seen without direct sunlight. We found that including this term in our model improves results and show its impact in Figure 5.15.

Figure 5.16 presents the results on the *castle* dataset. The base of the right tower is occluded by a tree and was therefore excluded from the object mask. The scanline through the towers shows a slightly flattened cylindrical shape on the left. The tower on the right shows a sharp corner with the x-component jumping from below -0.4 to about 0.4. This corresponds to an angle of more than 72° and thus differs about 15°

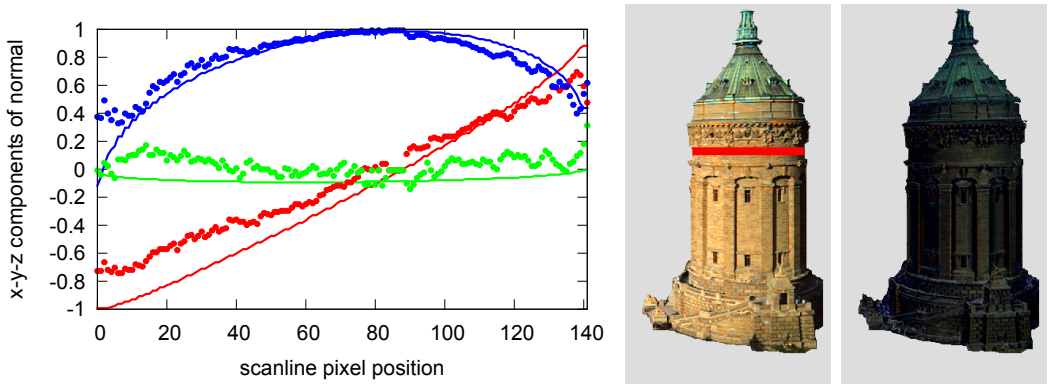


Figure 5.14: *Left*: Scanline through a cylindrical section of the *tower* showing the x (red), y (green), and z (blue) components of the normal vectors. Solid lines show normals from a reference cylinder, and dots show the reconstruction of our algorithm. *Center*: The region used for the scanline marked in red. *Right*: Recovered contribution of the skylight.

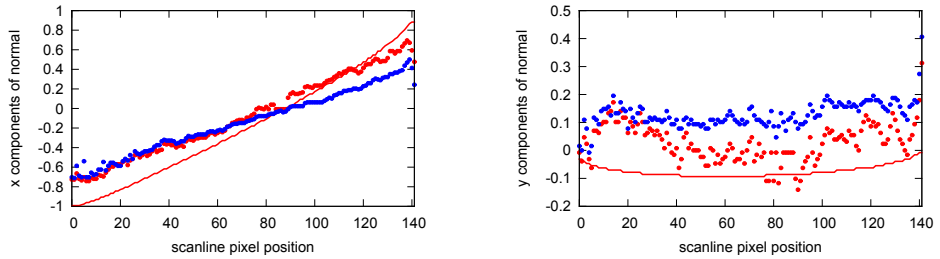


Figure 5.15: Impact of the sky contribution for the same scanline as marked in Figure 5.14. We plot the x (*left*) and y (*right*) components of the reference normals (red line), the final reconstruction with sky term (red dots) and without it (blue dots).

from the actual corner angle. Overall, the general directions are recovered correctly. We also notice that the material maps show a clear separation. This is in contrast to the other datasets where we often observe mixtures of materials. The separation decouples the pixels on the left side of the tower from the right side and leads to an ambiguity between the material and the sky term, which cannot be resolved by the optimization. The desired image intensities can either be achieved with a small sky term and a bright material or, as can be seen on the left side of the tower, with a material that is too dark and a bright sky term.

The last dataset is a large metal *sphere* that houses a roller coaster. It is especially interesting because of its highly specular surface. Figure 5.17 shows the reconstruction for two materials. The silver gray color and the specularity are captured correctly. We also observe that the recovered normals match those of an ideal sphere—ignoring the triangular facets. The noticeable errors in the normal map stem from a large ribbon that was draped around the sphere for several weeks before Christmas. These images violate the assumption of a static scene and were not detected as outliers during image selection.

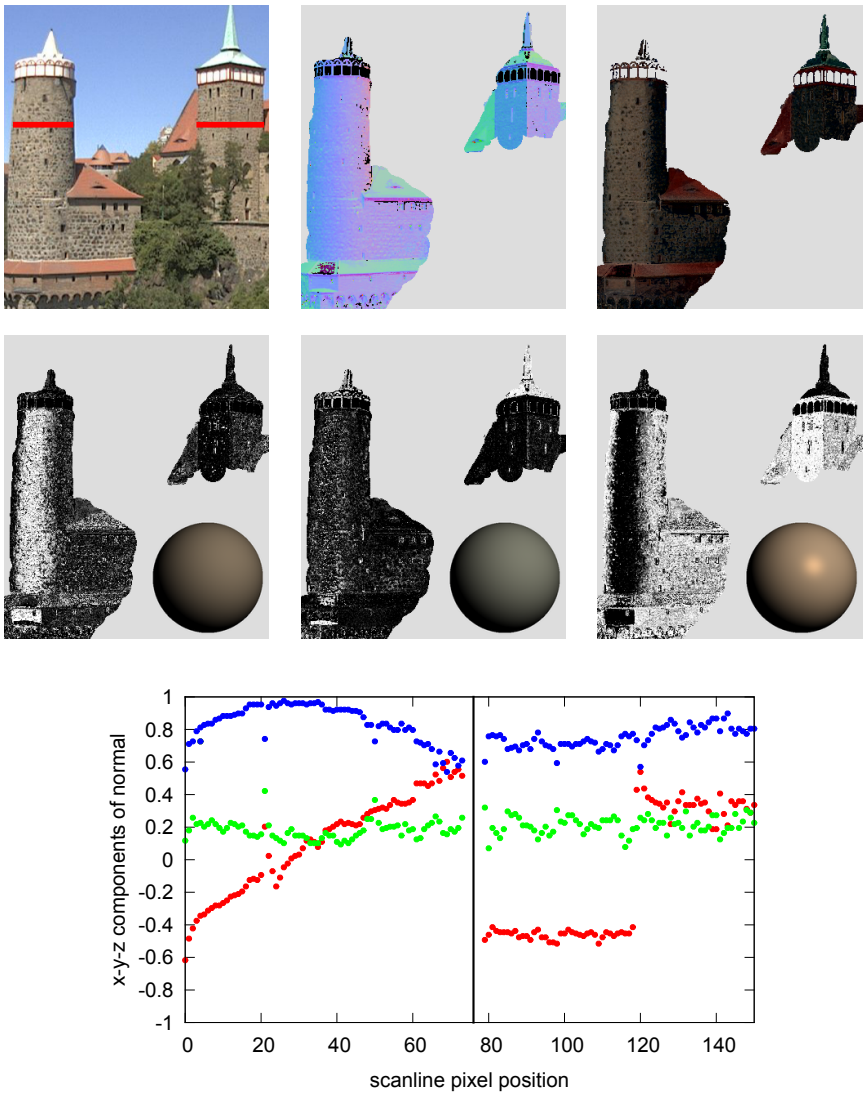


Figure 5.16: Results for the *castle* dataset [Weba]. *Top, left to right*: One of the input images with the scanline marked in red, the final normal map, and the estimated sky contribution. *Middle*: The reconstructed materials and mixture maps. *Bottom*: The x (red), y (green), and z (blue) components of the normal vectors along the scanline shown above.

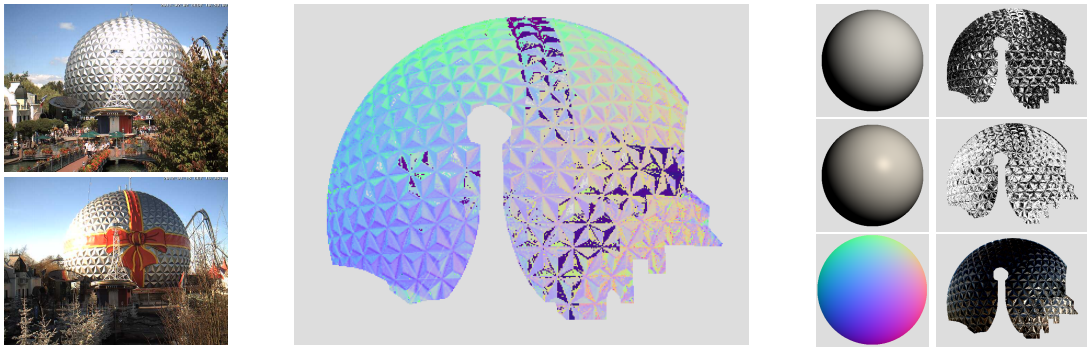


Figure 5.17: Results for the *sphere* dataset [Webc]. *Left*: Two input images. One shows a ribbon that was draped around the sphere for several weeks. *Middle*: The recovered normals match those of the ideal sphere shown to the right (bottom row). *Right*: The recovered materials and mixture maps (top rows). Normals of an ideal sphere and the reconstructed sky term (bottom row).

5.7 Discussion

In this chapter, we have shown how a model of outdoor illumination makes the problem of normal and reflectance recovery tractable even on Internet data. A very important insight also lies in the image selection which partially answers the question of what has to be adapted in existing techniques. Defining a suitable measure for "good" images in a specific reconstruction context is one of the key components when attacking these tremendous amounts of uncontrolled data. We have presented a working example for outdoor scenes based on heuristics. Developing new selection schemes for other illumination models or based on more explicit prior knowledge is a very promising route to lead photometric techniques out of the laboratory setting.

Another aspect of answering our motivating question for this chapter is the combination of several calibration steps. We could draw on recent advances in radiometric and geometric camera calibration techniques. These had never been integrated into a complete reconstruction pipeline. To obtain the light source positions, we exploit our outdoor illumination model and the fact that the position of the Sun can be calculated for any given time and location on Earth. Finally, we do not need to know the relative light intensities beforehand because they are modeled as degrees of freedom and estimated during the optimization. So far, typical pipelines, *e.g.* [Goldman05], have considered only controlled settings for preprocessing, calibration, and the actual reconstruction.

The approach is based on an energy minimization which is closely related to the image formation model and vice-versa. For example, factoring $I_{\mathrm{Sky},i,p,c} = l_{i,c} \cdot S_{p,c}$ into per-image and per-pixel components is largely motivated by the separate optimization stages and reduces the degrees of freedom. If the presented algorithm is applied in a different context, *e.g.* indoor webcams, the model and the optimization will have to be adapted. The challenge is to find a model with sufficient degrees of freedom to operate on general conditions but also sufficient constraints to obtain the desired solution. Apart from that, the fundamental ideas of using image selection and inverse rendering can easily be transferred.

The careful selection of model parameters is also the reason why we chose a rather

Figure 5.18: Relighting without intra-object shadows. *Left*: A webcam image that was not used as input to our reconstruction pipeline. *Center*: A re-rendered model of the *tower* based on our reconstruction with the same light direction—relative intensities were adjusted manually—inserted into the scene. *Right*: The rendering without scene context.

simple sky model. It is of course possible to replace the assumption of a uniform sky with a more sophisticated luminance distribution, such as the CIE sky models [Darula02], and might be interesting for future work. However, given the dominance of the sun intensity over the sky contribution, it is not clear whether a more detailed sky model would yield much better results. It also has to be clarified whether the optimization would still converge to the desired minimum given the additional parameters.

At the current state, we can certainly not reach the quality of techniques that assume known reflectance and only recover the shape of an object or any technique operating on controlled data. But the evaluation shows that the results are respectable given the type of input and the difficulty of *jointly* recovering shape and reflectance. Once the source of the tendency to flatten the normals is found—which might be caused by an inaccurate radiometric calibration, a wrong local optimum, interreflections, *etc.*—even better results should be possible. The current accuracy of normals and materials is already sufficient for one of the goals that motivate this thesis: simulating novel impressions of a scene. Figure 5.18 shows an image that was not part of the input and a rendering of our scene representation for the corresponding sun position. It is hard to tell which was the rendering even though small differences in shadow and color are present.

Future Work: As this was just a very first but important step towards photometric reconstructions on uncontrolled data, there remain several topics for future work. Similar to almost all techniques in this area, we have not considered the influence of interreflections and indirect lighting—apart from the light scattered in the atmosphere and making up the sky distribution. Especially the ground plane and nearby buildings can cause these effects and to handle them exactly would require at least a 3D model of the scene. Finding better ways to estimate global illumination effects from (outdoor) image sequences would immediately yield better reconstructions, as shown by Nayar *et al.* [Nayar90] in a controlled setting.

To improve the calibration, it would be beneficial to study the errors arising in the various steps and their possible accumulation. The radiometric calibration currently relies on diffuse surface points. While Kim *et al.* [Kim08a] show how to filter for

those, it would be preferable to develop ways that are independent of this assumption. Including the coefficients τ_j of the response curve in the overall optimization, as done by Abrams *et al.* [Abrams12] for diffuse scenes, might lead to instabilities in our formulation. Furthermore, the response has already to be known in the pose calibration step which also relies on linear intensities.

Another interesting aspect would be to incorporate temporal changes of the scene. The *sphere* dataset shows what happens if the static scene assumption is violated. In theory, webcams would be well suited to study weather effects on reflectance, *e.g.* wet on one day and dry on another. If these could be extracted automatically, a collection of more realistic appearances could be studied similar to the MERL database [Matusik03] for BRDFs or the advances by Bell *et al.* [Bell13].

Chapter 6

Fusing Multi-View Stereo and Photometric Stereo

So far, we have seen several contributions that were related to a calibrated photometric stereo setup. The discussion in Chapter 4 shows that it is definitely possible to apply the calibration of cameras and light sources as a preprocessing step. Extending on this, the pipeline in Chapter 5 demonstrates that a calibration is even possible on very uncontrolled Internet data. However, we have also seen that calibration needs a significant effort and can be tedious to implement. In addition, it is not perfect, and if several calibration steps are combined into a whole pipeline, this becomes even more of a problem as errors accumulate.

In this chapter, we are going to study the question whether it is possible to perform shape reconstruction from images without relying on calibrated lights or cameras with linear response. Any photometric stereo technique that is able to operate without a calibration also has a much better chance to be accepted as a tool outside of the research community. While there exists a whole body of literature on uncalibrated photometric stereo, these works all rely on Lambertian surfaces or certain assumptions on the illumination.

It turns out that the combination of two basic concepts is sufficient to achieve photometric reconstructions without any calibration—apart from the geometric camera parameters. The first one is *orientation consistency* and the second one is a *matching* of observed intensities against a known reference. In addition, basing our approach on orientation consistency circumvents an explicit modeling of reflectance. However, such an example-based approach as proposed by Hertzmann and Seitz [Hertzmann03,Hertzmann05] is only rarely considered because it involves placing a reference object—typically a sphere—in the scene. The downside of this approach becomes even more apparent if we return to one of the main topics of this thesis: *Internet data*. Ideally, we not only want to develop an algorithm that works without calibration, but that is also able to cope with input collected from the Internet. Of course, uncontrolled images or videos do not contain a reference object, and we cannot place anything in the scene.

Still, the generality of example-based photometric stereo is intriguing, and it allows us to reduce the challenge of surface reconstruction without calibration to the question whether it is possible to obtain suitable reference data solely from observations of the scene itself. Our key insight is that the geometry of many objects can be reconstructed

at least partially using multi-view stereo. State-of-the-art multi-view stereo methods are also sufficiently robust to work without a true radiometric calibration and can be applied to Internet images. We can thus replace the dedicated reference object with the requirement of additional images taken from several view-points from which we obtain a partial reconstruction. The advantage is that such a reference is intrinsic to the scene and can be recovered from images alone.

Moreover, orientation consistency is independent of the camera response if both the reference and target are observed in the same image. In practice, however, we want to allow for albedo changes and match intensity profiles even if they differ by a factor. This is not strictly correct for non-linear response curves because a scalar factor in scene luminance does not necessarily cause a proportional relation of image intensities. We did not find this theoretical discrepancy to produce many false matches and compute all results on non-linear images. In contrast to traditional approaches that compute the normal from image intensities, we only use the intensity values for matching which is less affected by the radiometric calibration.

Relying on another technique, such as multi-view stereo, raises the question why a photometric reconstruction is necessary at all. The rationale is that both approaches have different properties and advantages which are complementary to each other. Stereo often performs poorly on surfaces that exhibit only little texture variation. Additionally, any triangulation-based method has only a limited resolution in depth, which leads to high frequency noise in the normals. On the other hand, it provides a good estimate of the absolute geometry and not only recovers surface orientation as photometric stereo does. Photometric techniques work well even in uniform regions but only reconstruct a normal field, which has to be integrated in order to obtain the final surface.

Several techniques [Nehab05, Okatani12, Zhang12] have been proposed to combine positional measurements, *e.g.* from stereo, structured light, or laser scanners, with surface orientation. In our case, information from multi-view stereo not only enters the integration step but already influences the reconstruction of the normal field. This poses some additional challenges which we have to consider. First, the reconstructed reference geometry has to be as accurate as possible because it is the basis of all other computations. Second, the range of normal directions present in the reference geometry should cover a sufficiently large portion of the directions present in the complete scene. If only surfaces pointing downwards were part of the reference, the intensities observed for surfaces oriented upwards would never yield a correct match.

The design choices we make yield an algorithm that in principle works on very general input data. It is important to study how well these theoretical considerations transfer in practice. Furthermore, the influence of certain parameters, such as the matching score used to assess similarity, needs to be clarified. We therefore back our presentation by an extensive evaluation which we perform on controlled data as well as on Internet images.

6.1 Problem Statement and Overview

Our goal is to recover the surface of an object of unknown reflectance as accurately as possible from two sets of input images I^{MVS} and I^{PS}. The images $I_1, \ldots, I_M \in I^{\mathrm{PS}}$ show the object from a fixed camera position under unknown, distant, varying illumi-

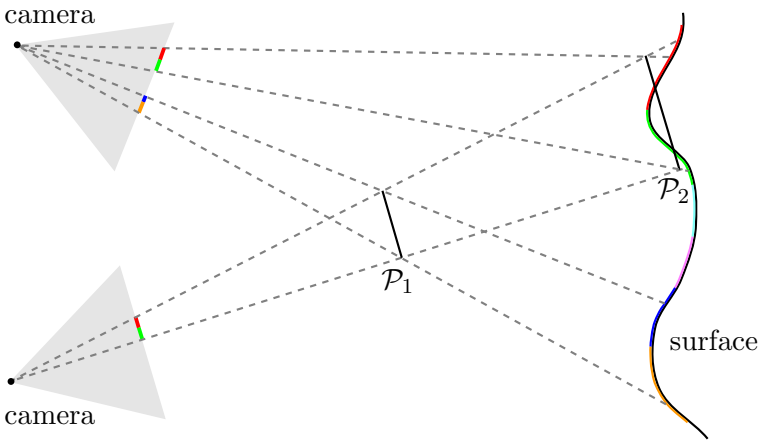


Figure 6.1: Patch-based stereo. A hypothetical patch $\mathcal{P}_1$ at a wrong position leads to different color values in both cameras. At the correct position $\mathcal{P}_2$, the projection yields a red and a green pixel in both cameras.

nation. We make the common assumption of an orthographic camera and represent the final surface as a height field $Z : \{1, \ldots, W\} \times \{1, \ldots, H\} \to \mathbb{R}$ defined over that camera. In contrast, the images in I^{MVS} are taken from different view points. They should provide sufficient parallax and lighting suitable for multi-view stereo reconstruction.

Our approach subdivides this general problem into three parts. We first reconstruct a partial geometric model that serves as scene intrinsic reference geometry. Using the reference geometry, we then aim at creating a complete normal map $N : \{1, \ldots, W\} \times \{1, \ldots, H\} \to \mathcal{S}$. We finally reconstruct the scene geometry by integrating the resulting normal field while taking the reconstructed reference geometry into account.

6.2 Scene Intrinsic Reference Geometry

In this first part, our goal is to create a geometric model of the object using multi-view stereo reconstruction techniques. While this kind of technique is really powerful, it often suffers from high frequency noise and mismatches in regions with only little texture variation. Thus, it does not directly solve our overall goal. But the approximate geometry can serve as an adequate starting point on which we will base our photometric reconstruction.

Multi-view stereo reconstruction works by projecting small (possibly oriented) 3D patches into neighboring camera views, *e.g.* [Furukawa10b]. For a hypothesized patch, the image intensities in all projections are compared using metrics such as "normalized cross correlation" or the "sum of squared differences" and aggregated into a similarity score. Based on this score, a hypothesis is either retained or discarded. The fundamental principle is that a patch at the 3D position of the surface will look similar in all images, whereas the cameras will observe different parts of the scene if the patch has a wrong position. Figure 6.1 illustrates this concept.

To first obtain the necessary projection matrices, we apply a robust structure from

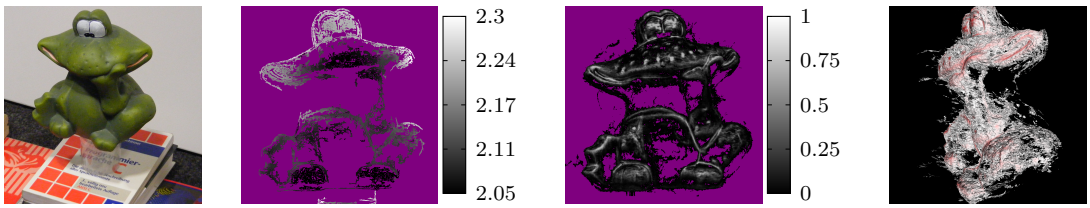


Figure 6.2: Example for the scene intrinsic reference geometry. *From left to right:* One of the input images for multi-view stereo. The reconstructed depth map for the photometric stereo view I_1. The confidence values of the global model merged from all depth maps using VRIP. The global model rendered from a different view (high confidence in *red*, low confidence in *gray*) shows lots of outliers and noise.

motion system as described by Snavely *et al.* [Snavely06] and sketched in Section 4.1. Adding one of the images in I^{PS} to all images in I^{MVS} during this step registers both sets into a common coordinate system. Note that this step assumes a perspective image formation model, whereas orientation consistency relies on an orthographic camera in theory. We keep the impact of this contradiction small by placing the camera far away and zooming in on the object.

Once the camera parameters are known, we run an existing multi-view stereo algorithm on all images in I^{MVS}. There is a large body of existing work on multi-view stereo reconstruction—see Seitz *et al.* [Seitz06] and the accompanying web page—and our proposed technique can be based on any of them. However, if the first step in our pipeline relied on a controlled capture setup, we could never hope to obtain any results on Internet data. Therefore, we select the method by Goesele *et al.* [Goesele07] which has a strong focus on robustness and very general input images. Furthermore, it is publicly available and easy to integrate into our framework.

This method reconstructs individual, incomplete depth maps using a region-growing approach. Figure 6.2 shows an exemplary input image and the corresponding depth map. We observe the afore-mentioned problem of missing reconstructions in homogenous regions like the belly. To exploit the redundancy present in depth maps from multiple views, we merge the reconstructions from all cameras into a combined triangular geometry model using *volumetric range image processing* (VRIP) proposed by Curless and Levoy [Curless96]. This slightly reduces noise and removes outliers, but the result still contains holes and spurious geometry because each of the individual reconstructions failed in these regions. Accordingly, VRIP assigns a low confidence to these vertices as shown in Figure 6.2.

However, some parts could be recovered quite well—especially at geometry boundaries and edges. We will base the next steps only on the reliable parts and throw away vertices with a confidence below τ_{conf}. The remaining geometry serves as a reference obtained from the target object itself and removes the need for an explicit reference object in the scene. Finally, we compute per-vertex normals for the reference geometry from surrounding face normals using area-weighted averaging.

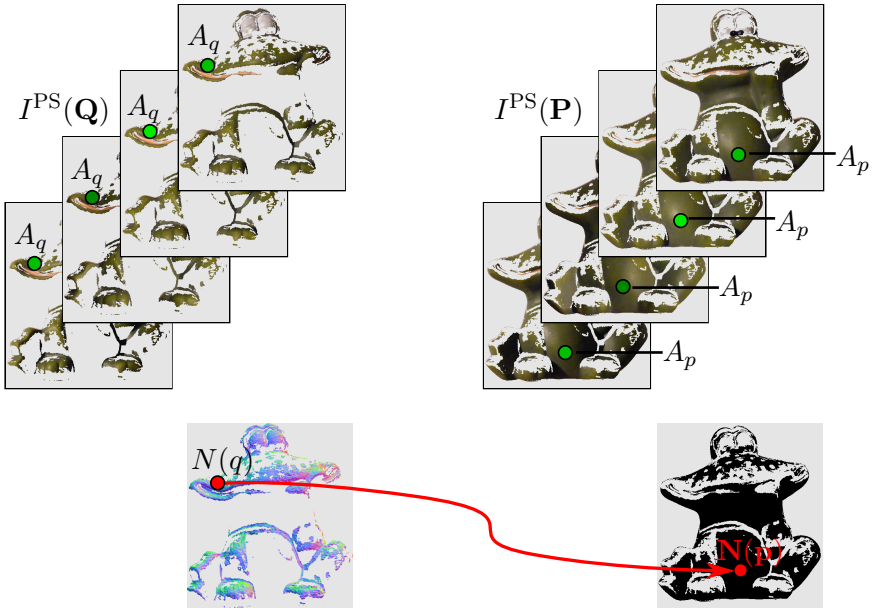


Figure 6.3: Matching and normal transfer example. *Top*: Intensities in the stack (shaded, green dots) form appearance profiles A_p, A_q which are matched between the scene intrinsic reference geometry $\mathbf{Q}$ (left) and the missing reconstruction $\mathbf{P}$ (right). *Bottom*: At matching positions, normals are transferred from the known regions.

6.3 Appearance-Based Normal Transfer

In this section, we describe the details of the photometric reconstruction part of our example-free approach. The key component is a matching step that selects similar appearance profiles. *Orientation consistency* then tells us that the corresponding normals are the same. In contrast to the setting of Hertzmann and Seitz [Hertzmann05], the scene intrinsic reference geometry is not a noise-free and complete reference object. We therefore introduce an orientation consistency based averaging and an adapted normal transfer approach to achieve improved reconstructions.

6.3.1 Matching

We manually segment the target object from the background in the first image I_1 of the photometric stack I^{PS} and then project the reference geometry into the corresponding view. All pixels in the mask are classified into those covered by the reference geometry $\mathbf{Q}$ and those for which no reconstruction is available $\mathbf{P}$. Figure 6.3 illustrates these sets for the *frog* dataset. For each pixel $q \in \mathbf{Q}$, we know its normal $N(q)$ from the rendered reference geometry. We furthermore define the *appearance profile* for each point in $\mathbf{P}$ and $\mathbf{Q}$ which is formed by all the color values for a particular pixel location in the image stack I^{PS}:

$$A_p = (A_{p,R}, A_{p,G}, A_{p,B}), \quad A_{p,c} = (I_{1,p,c}, \ldots, I_{M,p,c})^T, \quad c \in \{R, G, B\}. \tag{6.1}$$

The core of geometry completion is the appropriate transfer of normals derived from the scene intrinsic reference geometry to positions where reconstruction is missing. We therefore need to define a measure d of dissimilarity between two appearance

profiles. For every $p \in \mathbf{P}$, we then select the $q \in \mathbf{Q}$ for which $d(A_p, A_q)$ is minimal and transfer the normal $N(q)$ to $N(p)$. The process is illustrated in Figure 6.3.

In theory, it would be sufficient if that measure fulfilled

$$d(A_p, A_q) = 0 \quad \Longleftrightarrow \quad A_p = A_q. \tag{6.2}$$

In practice, we will rarely observe pairs that match exactly, *i.e.* $A_p = A_q$. The behavior of d for inexact matches $A_p \approx A_q$ is thus quite interesting. For example, the impact of intensity outliers on the matching result will differ depending on the choice of d. Another aspect to consider is whether the error measure has additional structure, *e.g.* if it is a metric, the triangle inequality can speed up the search for nearest neighbors.

In the following subsections, we will briefly present several error measures that we considered. They are all defined as the sum of per-color channel scores d_c:

$$d(A_p, A_q) = \sum_{c \in \{\mathrm{R,G,B}\}} d_c(A_{p,c}, A_{q,c}), \tag{6.3}$$

which is a metric if d_c is a metric.

Material Mixing

Hertzmann and Seitz [Hertzmann05] present a matching score that takes multiple reference objects into account. For each candidate p, they determine optimal mixing coefficients with all reference materials. Assuming a linear camera response, these coefficients can be interpreted as albedo differences. The mixing thus allows to match between reference objects with different color and even different reflectance properties. For a non-linear response curve, the interpretation in terms of albedo is no longer strictly correct but it is still possible to speak of mixtures of "materials"—which now incorporate the response function. Hertmann and Seitz mention that in practice the matching works well even for non-linear responses. In our case, we only have the reference geometry available and know nothing about reflectance clusters. But we would still like to benefit from the invariance to deviations up to a constant factor.

For a single reference, the score simplifies to

$$d_c(A_{p,c}, A_{q,c}) = \|m_{p,c} A_{q,c} - A_{p,c}\|_2^2 \tag{6.4}$$

where $m_{p,c}$ is chosen as the optimal per color channel material coefficient. This requires solving a simple least squares problem for each evaluation of d_c:

$$0 = \partial_z \|A_{q,c} z - A_{p,c}\|_2^2 = -A_{q,c}^\mathsf{T} A_{p,c} + A_{q,c}^\mathsf{T} A_{q,c} z \Rightarrow m_{p,c} = \frac{A_{q,c}^\mathsf{T} A_{p,c}}{A_{q,c}^\mathsf{T} A_{q,c}}. \tag{6.5}$$

In geometric terms, we find the orthogonal distance of $A_{p,c}$ to the ray defined by $A_{q,c}/\|A_{q,c}\|$. This is not a symmetric operation as illustrated in Figure 6.4 and thus does not yield a true metric.

Norm Induced Metrics

The appearance profiles in a single color channel can be interpreted as elements of a vector space $\mathbb{R}^M$. To again allow for a constant factor between the target and reference profile, we define an equivalence relation

$$A_{p,c} \equiv A_{q,c} \quad :\Longleftrightarrow \quad \exists\, \alpha_c \in \mathbb{R}_+ : A_{p,c} = \alpha_c A_{q,c} \tag{6.6}$$

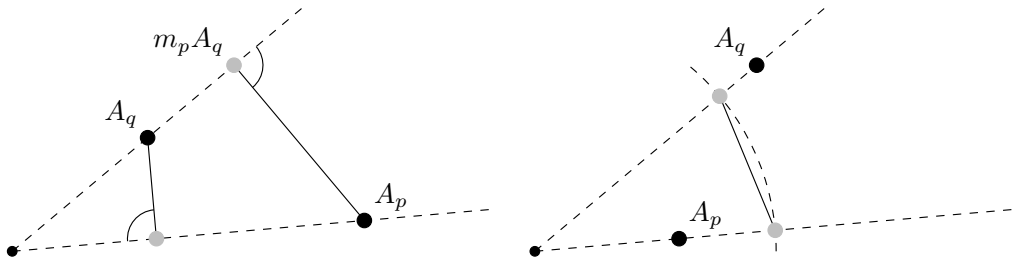


Figure 6.4: Geometric interpretation of matching scores. *Left*: Given A_p, we find the closest point on the ray from the origin to A_q and use its Euclidean distance. *Right*: Points are projected onto the unit sphere and then compared using a standard norm (here: L_2).

with normalized profiles $B_{p,c} = A_{p,c}/\|A_{p,c}\|$ as representatives of the equivalence classes $\lfloor B_{p,c} \rfloor$. We now redefine an appearance profile as one of these equivalence classes and consider the induced metric on the quotient space:

$$d_c(\lfloor B_{p,c} \rfloor, \lfloor B_{q,c} \rfloor) = \|B_{p,c} - B_{q,c}\|. \tag{6.7}$$

In Section 6.5.1, we will study the L_1 and L_2 norms, but the general procedure can be applied to any norm. The geometric interpretation illustrated in Figure 6.4 is that of replacing "points" with "lines through the origin" as in projective geometry.

Root Mean Square Error

We also study one score that does not explicitly allow for color differences: the root mean square error (RMSE) of two profiles. This allows us to easily incorporate a basic shadow handling. The image formation model at the heart of most photometric stereo techniques and also underlying the orientation consistency assumption does not account for occluders between the light source and a surface point. Therefore, cast shadows are typically considered as outliers. If we stack all appearance profiles into an *observation matrix*, this amounts to categorizing pairs (i, j) into shadow or non-shadow classes. Removing these entries from the observation matrix leads to the problem of missing data during matching.

We simply use a threshold τ_{sh} to identify shadows and define the set

$$\mathcal{I}_{p,c} := \{1 \leq i \leq M \mid I_{i,p,c} > \tau_{sh}\} \tag{6.8}$$

of non-shadowed images for each pixel. Our matching score for two profiles is then defined on the intersection of the two respective index sets:

$$d_c(A_{p,c}, A_{q,c}) = \Big(\frac{1}{|\mathcal{I}_{p,c} \cap \mathcal{I}_{q,c}|} \sum_{i \in \mathcal{I}_{p,c} \cap \mathcal{I}_{q,c}} (I_{i,p,c} - I_{i,q,c})^2 \Big)^{1/2}. \tag{6.9}$$

6.3.2 Averaging to Counter Noise

Figure 6.5 shows the normals obtained from VRIP which are contaminated by high frequency disturbances, even though we use a conservative confidence threshold. A standard approach to reduce these effects is mesh smoothing. We iteratively update

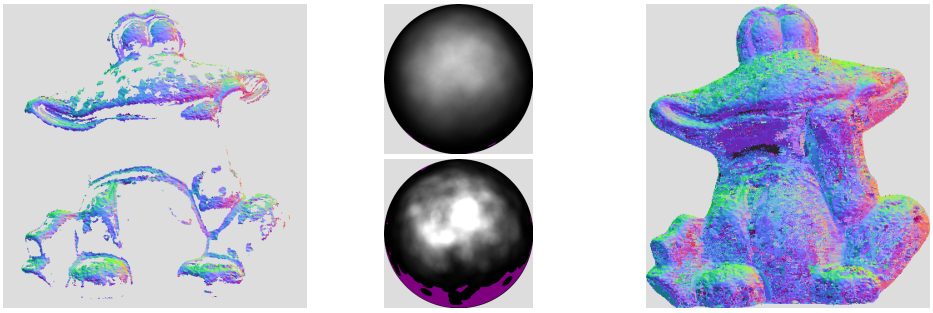


Figure 6.5: Challenging input data. *Left:* Normals of the VRIP reconstruction show high frequency disturbances. *Middle:* Too much smoothing removes orientations from the reference that cannot be recovered afterwards. The visualization shows the distribution of directions present in the VRIP model (*top*) and after heavy smoothing (*bottom*) with equal color mapping (missing entries are marked in *purple*). *Right:* Transferred normals from **Q** to **P** are very noisy.

a normal by interpolating between its current direction and the average normal computed from neighboring vertices $\mathcal{J}$:

$$\tilde{n}^{i+1} = (1-\lambda)\cdot n^i + \frac{\lambda}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} n^i_j, \qquad n^{i+1} = \tilde{n}^{i+1}/\|\tilde{n}^{i+1}\|. \tag{6.10}$$

Smoothing leads, however, to a shrinked set of normal directions and also reduces the quality of those normals that have been reconstructed correctly. The visualization in Figure 6.5 is a hemispherical histogram where each normal contributes a fixed amount of "mass". Since we rely on a transfer of normals, we can only recover those normal directions that are already present in the reference geometry. We are thus faced by a trade-off between a smooth reconstruction and a detailed one. We found heavy smoothing to introduce a strong bias towards a few favored directions and use ten iterations ($\lambda = 0.05$) for our experiments.

Figure 6.5 shows the result of transferring these normals in **Q** to **P** based on the score in Equation (6.4). We observe that the normals in the transferred regions **P** are much noisier than the initial ones in **Q** and are in fact completely useless. This could be due to several reasons: First, some normal directions are not represented in the reference geometry, *e.g.* most of the downward pointing normals on the *frog*'s neck area. Second, the matching according to the orientation-consistency cue might fail if appearances differ too much from the reference. Third, even if the matching is correct, the reference geometry still contains erroneous normal information.

On the other hand, if we plot the locations and directions of the s best matches for a single pixel in Figure 6.6 ($s = 50$), we notice that most normals are clustered around an average direction. We therefore propose to not only use the normal corresponding to the best-matching intensity profile but to compute an average normal from the s best matches. This reduces the impact of wrong matches and erroneous normals as shown for $s = 50$ and $s = 100$ in Figure 6.6. We found $s = 100$ to provide only small improvements and use $s = 50$ for all datasets.

An additional benefit of the averaging step is that the transferred normals are no longer limited strictly to the set of reference normals. For example, if the reference normals were either looking to the left or right, we could interpolate normals facing

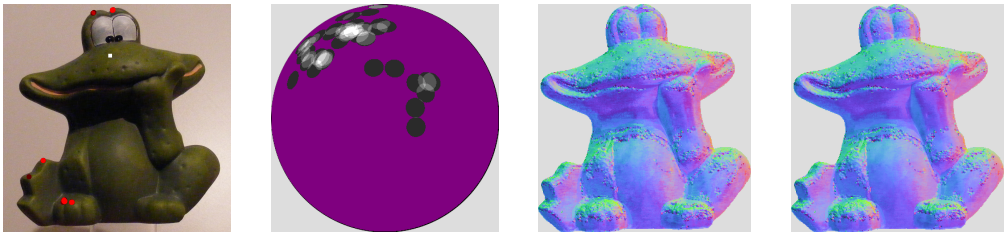


Figure 6.6: Averaging several matches. *From left to right:* For a pixel (*white*) in **P**, the position of the 50 best matches in **Q** is marked in *dark red* (most matches overlap and lead to bright red spots). The normals of the best matches are visualized as cone intersections with a hemisphere. Averaging the 50 best normals during transfer leads to a much smoother result. Increasing the number to 100 provides no benefit.

to the front. In practice, the correctness of such an interpolation does of course also depend on the appearance profiles and on the distribution of the best matching directions, but it still adds some flexibility to the transfer step. Note that it will not fix the case of normal directions outside the convex hull of normals observed in the reference geometry, but it may at least assign a nearby direction inside the convex hull.

Now, the result in **P**, shown in Figure 6.6, appears less noisy than the original parts in **Q**. As we discussed, overly smoothing the latter beforehand is not an option since it removes normal directions and introduces a bias towards often occurring directions. To get a consistent result, we instead propose to apply the averaging just introduced to **Q** also. More formally, we adapt the matching to transfer normals from **Q** to both **P** *and* **Q**. The difference compared to regular mesh smoothing is that normals are combined based on closeness in "appearance space", and not because their vertices are close in terms of surface distance. The latter is, however, included in the former, as demonstrated by the many overlapping matches in Figure 6.6 (left). We return to the effect of appearance based averaging on **Q** in Section 6.5.1.

6.4 Surface Reconstruction

Most photometric stereo methods recover the gradient of a surface, *i.e.* its orientation. On the other hand, it is often desirable to reconstruct the actual surface, and not just its derivative. If the object is represented as a height field $(u, v, Z(u, v))$ over the image plane, its gradient field is defined as

$$\nabla Z : W \times H \to \mathbb{R}^2, (u, v) \mapsto \begin{pmatrix} \frac{\partial Z(u,v)}{\partial u} \\ \frac{\partial Z(u,v)}{\partial v} \end{pmatrix} = \begin{pmatrix} Z_u(u, v) \\ Z_v(u, v) \end{pmatrix}. \tag{6.11}$$

What we obtain from photometric stereo is a vector field

$$g : W \times H \to \mathbb{R}^2, (u, v) \mapsto \begin{pmatrix} p(u, v) \\ q(u, v) \end{pmatrix} \tag{6.12}$$

or

$$N : W \times H \to \mathcal{S}, (u, v) \mapsto \frac{1}{\sqrt{1 + p^2 + q^2}} \big(p(u, v), q(u, v), -1\big). \tag{6.13}$$

One strategy to recover the surface is to "integrate" the gradient field. Let the photometric stereo reconstruction be perfect, *i.e.* $p = Z_u, q = Z_v$, and the gradient be differentiable. Then, the definition

$$\tilde{Z}(u,v) := \int_\gamma g(w) \cdot dw \tag{6.14}$$

is independent of the choice of γ (which is an arbitrary path from $(0,0)$ to (u,v)). Conversely, it holds that $\nabla\tilde{Z} = g$. Thus, the path integral in Equation (6.14) is one way to obtain a solution. It is, however, not unique and might differ from the true solution by an additive constant because $\nabla(\tilde{Z} + c) = \nabla\tilde{Z} = g$. Examples of this strategy can be found in [Wu88, Klette96].

In practice, the reconstruction is not perfect: the recovered gradients might differ from the true surface derivatives or they might not form an integrable vector field. It is therefore more common, *e.g.* [Durou09], to define the solution in a variational framework and minimize

$$E(\tilde{Z}) = \iint \big(\tilde{Z}_u(u,v) - p(u,v)\big)^2 + \big(\tilde{Z}_v(u,u) - q(u,v)\big)^2 d(u,v) \tag{6.15}$$

or similar error measures. Moreover, such a formulation can be incorporated directly into the reconstruction algorithm as shown by several shape from shading approaches, *e.g.* [Horn86b], that compare the reflectance map R against image intensities I:

$$E(\tilde{Z}) = \iint \big(I(u,v) - R(\tilde{Z}_u, \tilde{Z}_v)\big)^2 d(u,v). \tag{6.16}$$

To avoid outliers and discontinuities, it can be beneficial to multiply the gradient terms with spatially varying weights as for example shown by Agrawal *et al.* [Agrawal06].

Another approach to obtain absolute depth from gradients is to add additional information to guide the integration. This usually comes in the form of known depth values $\hat{Z}$ at sparse points or in the whole image area obtained through other techniques, *e.g.* laser scanning [Horovitz04]. Combining normals and depth in such a way is a key part in many of the multi-view photometric stereo methods discussed in Section 3.6. A possible extension of Equation (6.15) is, for example,

$$E(\tilde{Z}) = \iint \big(\tilde{Z}_u - p\big)^2 + \big(\tilde{Z}_v - q\big)^2 + \big(\hat{Z} - Z\big)^2 d(u,v). \tag{6.17}$$

We follow a similar approach as proposed by Nehab *et al.* [Nehab05]. They assume dense depth and normal maps Z, N. Because the former are corrupted by high frequency noise and the latter by low frequency bias, they first compute a new set of normals using information from both. In our case, depth information is only available in $\mathbf{Q}$, so that we have to omit this step. Nehab *et al.* replace the depth map Z with 3D positions $R = (Z \cdot r_u, Z \cdot r_v, Z)$ determined by their distance from the camera along the ray $r = (r_u, r_v, 1)$. Then, the discretized error function is defined as

$$E(\tilde{Z}) = \alpha \sum_j \|\hat{R} - \tilde{R}\|^2 + (1-\alpha) \sum_j \big[\langle N,\ T^u(\tilde{R})\rangle^2 + \langle N,\ T^v(\tilde{R})\rangle^2\big] \tag{6.18}$$

where we omitted the indices j for better readability. Instead of comparing the difference of optimized normals $\tilde{N}$ to initial normals N, they use the dot product between

the tangents to the optimized surface T^u, T^v and the given normal as an error metric. This is key for the high efficiency of the approach, because it allows to formulate the optimization as a least squares problem with a sparse matrix. The tangent T^u of $\tilde{R}$ is related to the gradient of $\tilde{Z}$ as

$$T^u = \frac{\partial \tilde{R}}{\partial u} = \frac{\partial \tilde{Z}}{\partial u} \cdot r + \tilde{Z} \cdot \frac{\partial r}{\partial u} = \Big(r_u \frac{\partial \tilde{Z}}{\partial u} + \tilde{Z} \frac{\partial r_u}{\partial u},\, r_v \frac{\partial \tilde{Z}}{\partial u},\, \frac{\partial \tilde{Z}}{\partial u} \Big). \tag{6.19}$$

If we stack all entries of the two-dimensional $\tilde{Z}$ into a single vector, we can formulate the approximation of $\partial \tilde{Z}$ with finite differences as a vector multiplication. For example, the basic central difference for the j-th entry yields

$$\frac{\partial \tilde{Z}}{\partial u}(j) = \begin{pmatrix} \cdots & -1 & 0 & 1 & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ \tilde{Z}_{j-1} \\ \tilde{Z}_j \\ \tilde{Z}_{j+1} \\ \vdots \end{pmatrix} = \underbrace{\big(-C(j-1) + C(j+1) \big)}_{=: D^u(j)} \tilde{Z} \tag{6.20}$$

where $C(j) = (\delta_{k,j})_{k=1}^M$ selects the j-th component of $\tilde{Z}$. Combining Equations (6.19) and (6.20), we obtain

$$\begin{aligned} \langle N_j,\, T^u(\tilde{R}_j) \rangle &= \frac{\partial \tilde{Z}}{\partial u}(j) \cdot \langle N_j,\, r_j \rangle + \tilde{Z}_j \cdot \big(N_{j,u} \frac{\partial r_{j,u}}{\partial u} \big) && (6.21) \\ &= \Big(\langle N_j,\, r_j \rangle D^u(j) + N_{j,u} \frac{\partial r_{j,u}}{\partial u} C(j) \Big) \cdot \tilde{Z} && (6.22) \\ &=: V^u(j) \cdot \tilde{Z} && (6.23) \end{aligned}$$

and a similar formulation for T^v. We stack these row vectors into matrices V^u, V^v and define a diagonal matrix $U := \operatorname{diag}(\|r_1\|, \ldots, \|r_M\|)$. Then, the error function transforms into

$$E(\tilde{Z}) = \left\| \begin{pmatrix} \alpha U \\ (1-\alpha) V^u \\ (1-\alpha) V^v \end{pmatrix} \cdot \tilde{Z} - \begin{pmatrix} \alpha U \hat{Z} \\ 0 \\ 0 \end{pmatrix} \right\|^2. \tag{6.24}$$

In contrast to Nehab *et al.* [Nehab05] and Joshi and Kriegman [Joshi07], we do not have positional constraints available at all pixels. We therefore remove rows from U that correspond to pixels in $\mathbf{P}$. The resulting matrix is still overdetermined because each pixel contributes at least two rows in V^u, V^v. Even for more complex finite difference schemes, the sum over each row is zero and thus $(1, \ldots, 1) \in \ker V^u, \ker V^v$. One can show that this vector spans the kernel—if there are no disconnected pixel regions—by an argumentation along the lines of $\partial Z(x,y) = 0, \forall (x,y) \Rightarrow Z(x,y) = \text{const}$. Accordingly, if U contains at least one row—which has exactly one non-zero entry—the overall matrix has full rank.

6.5 Evaluation

We have already shown a few results to give a better intuition about some of the design choices. We will now present a more thorough evaluation and conduct experiments

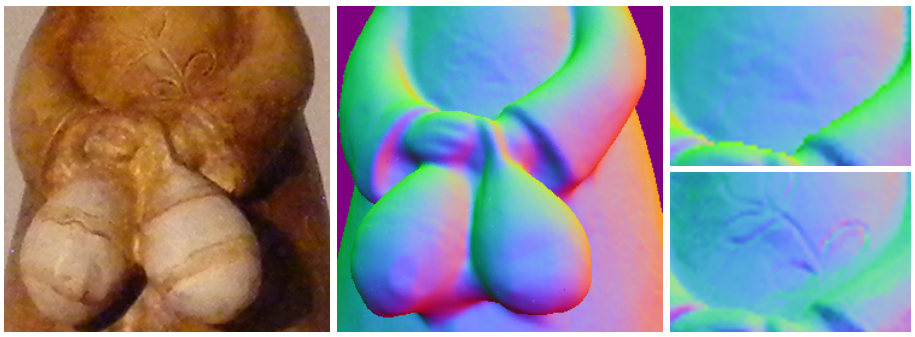


Figure 6.7: Quality of the ground truth. We show a closeup of an input image (*left*) and the ground truth normal map (*middle*) for the *bunny* dataset. The structured light scanner leads to vertical artifacts and a slight blurring of details in the normal map, but is still suitably accurate for comparison purposes. We also observe that photometric stereo normals (*right, bottom*) encode more details, but are slightly less pronounced in concave regions compared to the structured light reconstruction (*right, top*).

to clarify the performance of our approach. To this end, we captured datasets under controlled as well as uncontrolled conditions including Internet data. Furthermore, we test individual aspects of our approach with experiments on synthetic images that provide an idealized environment.

6.5.1 Lab-based Datasets

We begin with the evaluation on datasets captured in a laboratory setting. This allows us to acquire a ground truth and perform quantitative comparisons, which is much harder on Internet data. All images were shot with a consumer camera (Fuji FinePix S5700) and have been processed at the full resolution of 3072×2304. We do not apply any kind of gamma correction or inverse response curve, which other methods depend upon.

To obtain quantitative results, we acquired a ground truth mesh using a structured light scanner. The models show less deviations than the multi-view stereo reconstructions and better represent the true surface. They do, however, introduce slight, vertical step artifacts caused by the scanning procedure. Also, the structured light approach blurs some of the fine details as demonstrated in Figure 6.7. Furthermore, we automatically transform the ground truth mesh into the coordinate system of our reconstruction using the silhouette alignment technique proposed by Lensch *et al.* [Lensch00]. This leads to a very accurate alignment as we verified by projecting the mesh into the original images. Nevertheless, a small error remains that might bias the comparison. Still, the scanned meshes provide a suitable baseline to compare against, but we should keep these issues in mind when interpreting the results.

For each dataset, we captured $14 - 20$ images from a fixed camera position while manually moving a light source around the object. The position and intensity of the light are unknown. We took care to place the light several meters away to fulfill the distant lighting assumption. The camera was about 2 m away from the object. We additionally captured about 50 I^{MVS} images from various positions facing the front of the objects. These images also show a larger portion of the scene and contain other

Figure 6.8: Lab-based data. *Left:* Exemplary input images for the *frog*, *bunny*, and *bust* dataset. *Right:* Rendered normals of an ideal sphere to define the color mapping used for all normal maps.

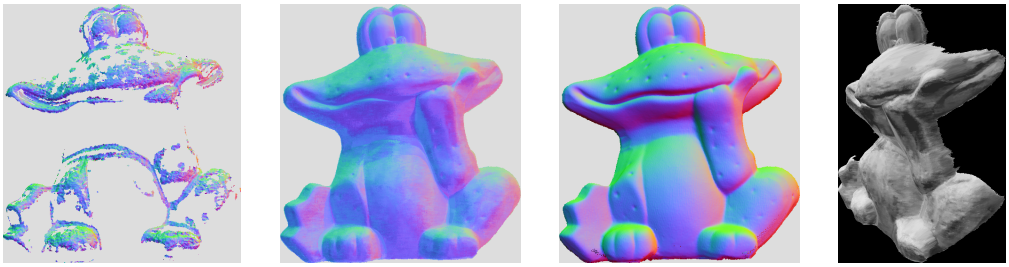
Dataset	Images	Normals	τ_{conf}	Coverage	Median Deviation	
				$\lvert\mathbf{Q}\rvert/\lvert\mathbf{Q}\cup\mathbf{P}\rvert$	Initial	Final
frog	14	700K	0.1	34%	17.2°	13.6°
bunny	15	300K	0.3	47%	12.1°	10.0°
bust	20	1700K	0.1	53%	15.8°	13.6°
tower	36	56K	0.3	58%	–	–
church	11	45K	0.3	84%	–	–

Table 6.1: Dataset overview. The first three datasets are captured under controlled, but unknown conditions. The last two rows use uncontrolled Internet images. We compute the angular deviation from the ground truth for all normals in the initial geometry on $\mathbf{Q}$ and in the final reconstruction on $\mathbf{Q}\cup\mathbf{P}$. We interpret deviations above 45° as outliers and take the median over the remaining set.

objects, such as the carpet, which exhibit sufficient features for the structure from motion step.

The first dataset is a painted clay *frog* of 25 cm height shown in Figure 6.8. Its reflectance is close to diffuse, and the surface shows large homogenous regions. The concavity at the neck is in shadow in several of the input images. The *bunny* in Figure 6.8 is a plastic figurine with shiny coating. It is about 20 cm tall and shows finer details than the *frog*. The final dataset in this section is a bronze *bust*. We use it to demonstrate the performance for challenging reflectance properties. Additionally, it exhibits complex surface structure. Table 6.1 summarizes some facts about each of these datasets.

Figures 6.9, 6.10, and 6.12 show the qualitative results for all controlled datasets computed with Equation (6.4). We render the integrated surface from a novel view point (large scale distortions are avoided due to the inclusion of reference geometry in the integration routine), but focus on the normal maps for the evaluation because their transfer is at the core of our technique. We observe that the reconstructed normals are quite plausible, even though the reference is rather incomplete. At the same time, we notice that our results (b) are less pronounced than the ground truth (c), *i.e.* the color-mapped normals are less saturated. This corresponds to a slightly flattened shape, which we attribute to the smoothing that is necessary because of the extremely noisy input data. The neck of the frog and the area around its right leg



(a) (b) (c) (d)

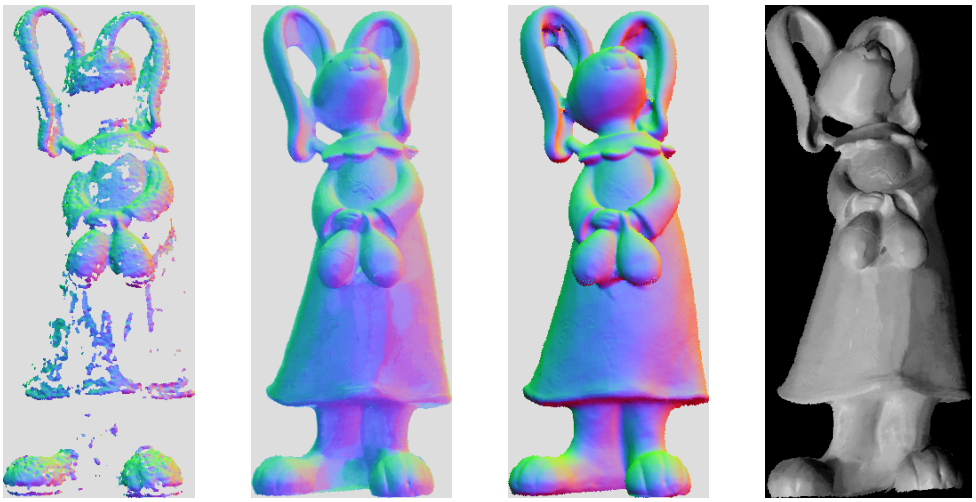
Figure 6.9: The *frog* dataset. a) The reference normals reconstructed from the scene itself. b) Our resulting normal map after matching and smoothing. c) The ground truth normals acquired from a structured light scanner. d) The integrated surface rendered from a novel view.

show some effects of shadows. This leads to consistent, but incorrect matches in these regions. The areas have homogenous normals as desired but deviate from the ground truth as a whole. Overall, the amount of detail is quite high as we see, for example, at the eyes of the *bust* or the carvings on the *bunny*.

For the quantitative comparison, we project the ground truth model with associated normals into the camera. We then compute the per-pixel angular deviation $\operatorname{acos}(\langle N,\ N^{\mathrm{GT}}\rangle)$ between the reconstructed normal N and the ground truth normal. First, we evaluate the performance of different matching scores as introduced in Section 6.3.1: the metrics induced by the L_1 and L_2 norm, the least squares error for the optimal material coefficient (*Coeff*), and the RMSE (with shadow threshold at 10 %). Figure 6.11 contains normalized histograms over the angular error for the *frog* and *bunny* datasets. We observe that there is hardly any difference in performance among the proposed scores. This can probably be attributed again to the averaging of the best matches which equalizes the results. We apply the "optimal coefficient" score in all subsequent results because it has the potential to be extended to several material clusters, which we would like to address in the future.

For the *bust*, we look at the effect of the appearance based averaging in $\mathbf{Q}$. Figure 6.12 shows a normal map without modifying normals in the reference area (a) and next to it our actual result (b). In both, the left side of the nose is not as strongly slanted as the true surface, but the smoothness of the final result better resembles the ground truth (d). To demonstrate that our technique works also with more general lighting, we captured two additional datasets for the *bust* where we used a studio light with and without a diffuser for illumination. Note that abandoning the point light source assumption present in many other approaches is a key issue when dealing with uncontrolled data. The closeups in Figure 6.13 demonstrate that the recovered normals change only marginally. We use the original dataset in the following quantitative comparisons.

If we assume that appearance changes slowly over the surface, then it is less likely for a point $p \in \mathbf{P}$ to find a good match than for a point in $\mathbf{Q}$. We therefore investigate the quantitative error with respect to its spatial distribution. Figure 6.14a shows histograms of the final reconstruction over the whole object similar to Figure 6.11. The other two plots (Figure 6.14b, c) contain histograms computed on $\mathbf{Q}$ and $\mathbf{P}$ separately.



(a) (b) (c) (d)

Figure 6.10: The *bunny* dataset. a) The reference normals reconstructed purely from images of the scene. b) Our resulting normal map plausibly fills in missing regions. c) The ground truth normals. d) A rotated view of the final surface.

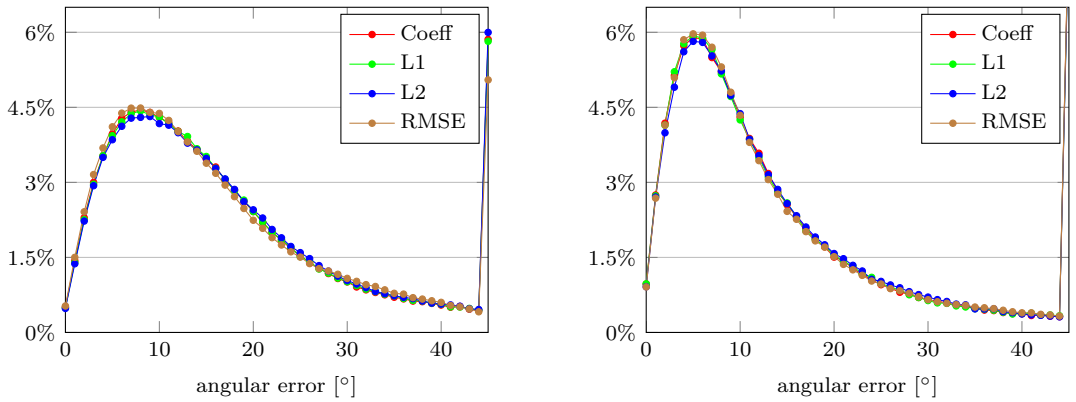
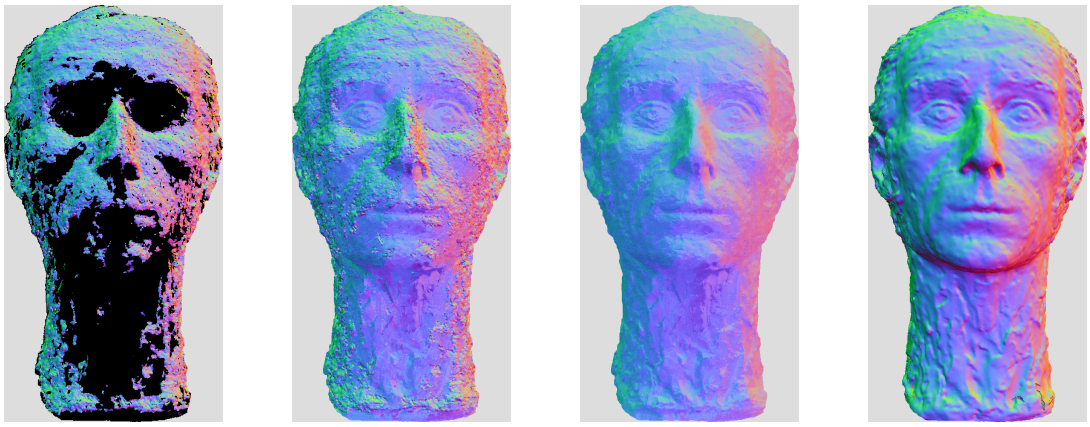


Figure 6.11: Histograms of angular deviation between the ground truth and the reconstructions from different matching scores. *Left*: The *frog* dataset. *Right*: The *bunny* dataset. Errors above 45° accumulate to 8.4 % (Coeff), 8.2 % (L_1), 8.2 % (L_2), and 8.4 % (RMSE).

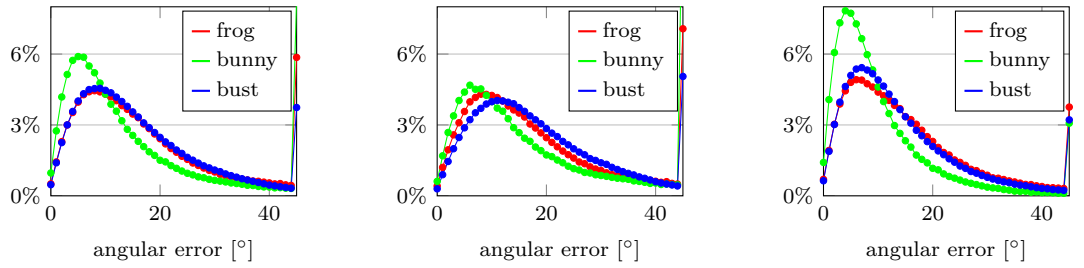


(a) (b) (c) (d)

Figure 6.12: The *bust* dataset. a) The incomplete reference normals. b) Normals transferred from **Q** to **P** are less noisy due to the averaging of best matches. c)Applying appearance-based averaging also to **Q** yields more consistent normals. d) The ground truth normals are more pronounced than both, the initial and final, normal maps.

(a) point light (b) spot light (c) diffuse light

Figure 6.13: General lighting conditions. a) Closeup of the reconstruction from Figure 6.12. The images were taken under a point light source. b) The results change only marginally if we use a focused spot light. c) Reconstruction from images taken with an extended light source.



(a) histogram on $\mathbf{Q} \cup \mathbf{P}$ (b) histogram on $\mathbf{P}$ (c) histogram on $\mathbf{Q}$

Figure 6.14: Angular deviation between the ground truth and the reconstructions for all datasets. Histograms are computed on $\mathbf{Q} \cup \mathbf{P}$ (a), $\mathbf{P}$ (b), and $\mathbf{Q}$ (c).

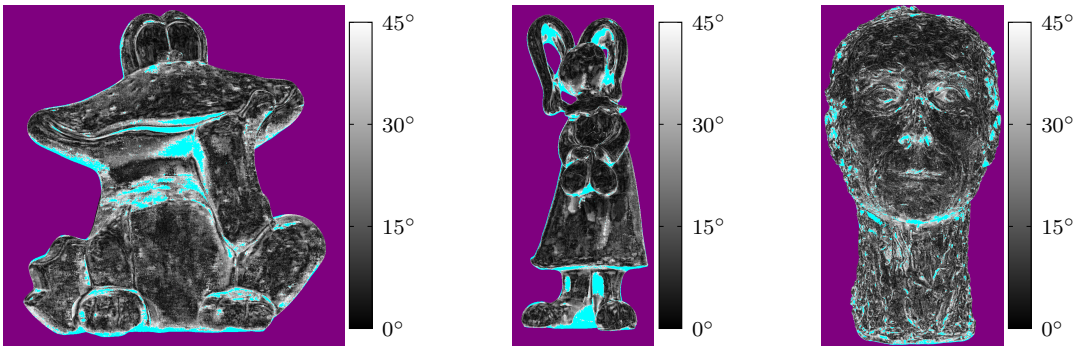
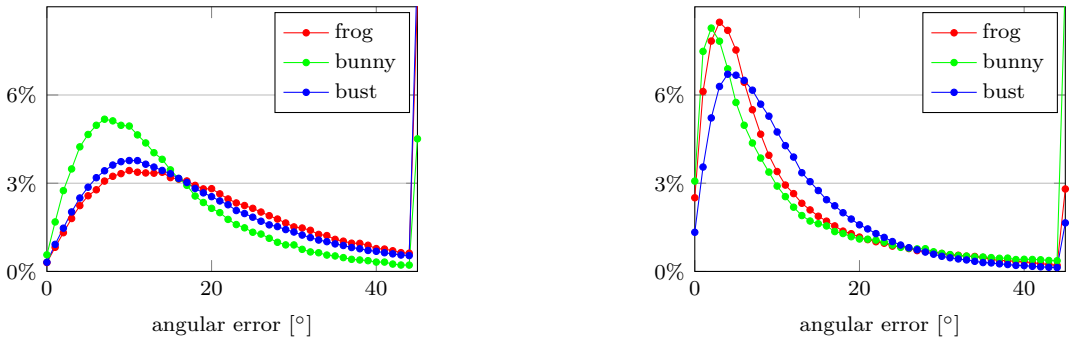


Figure 6.15: Spatial distribution of angular error. Deviations of more than 45° are marked in *cyan*. Most errors occur at depth discontinuities and in shadow regions.

We observe that, as expected, the performance in those regions that correspond to the reference is better than in those that have to be filled in. We also conclude that the overall result, Figure 6.14a, of the *bunny* is mostly due to its advantage in the reference area, Figure 6.14c, compared to the other datasets. A more detailed view of the error distribution is provided by Figure 6.15. The color-coded images show the angular error after applying our full pipeline. We note that the largest errors occur in concave regions affected by shadows. This is in line with our discussion about the qualitative results. We also observe many outliers at depth discontinuities, *e.g.* the skirt of the bunny. One explanation for these is a misalignment between our result and the ground truth. The other is that the reconstruction of surface patches observed under grazing angle is unstable in general.

We have argued several times that the low quality of the input geometry is one of the main challenges we have to face. To quantify this, Figure 6.16a shows the angular deviation of the initial reference normals from the ground truth. The histogram is shifted to the right, indicating greater deviations, compared to our final reconstruction (Figure 6.14c). Thus, our approach not only fills in missing normals but also improves those of the reference. Nevertheless, the errors in the input normals are rather large and we cannot expect an exact reconstruction. The dependence on input data becomes even more apparent if we imagine the case of multi-view stereo delivering perfect normals. We simulate this by supplying the ground truth normals in $\mathbf{Q}$ as reference



(a) error of initial normals

(b) error on **P** for ideal initialization

Figure 6.16: Dependence on input quality. a) Angular deviation of initial normals and ground truth on **Q**. b) Angular error on **P** if we had perfect input normals available on **Q**.

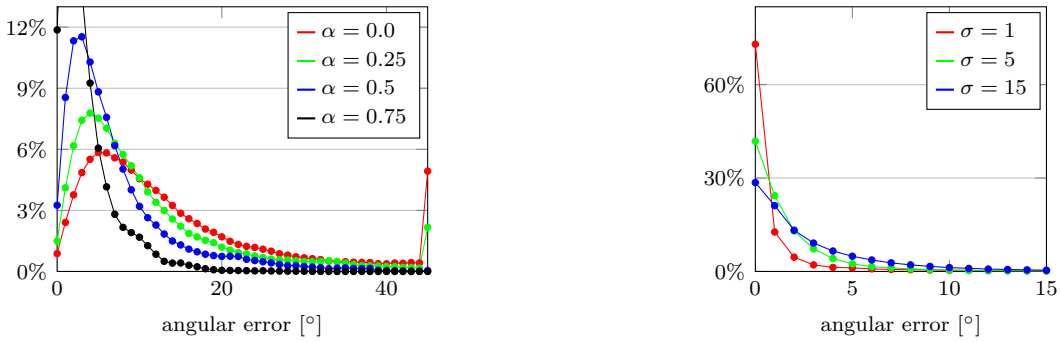
to our algorithm. The reconstruction results on **P** are shown in Figure 6.16b and are much better than those in Figure 6.14b.

6.5.2 Synthetic

The previous subsection has shown that we are limited by the reference normals which are far from perfect. To assess the impact of other imperfections, we use a synthetic dataset of linear, high dynamic range images. These are rendered from the ground truth 3D model of the *frog* with Lambertian reflectance. We use 20 different, ideal directional sources without any interreflections using the physically-based ray tracer published by Pharr and Humphries [Pharr12].

In Figure 6.16b, we based the reconstruction on real input images and optimal reference normals. In contrast, we now consider optimal images and imperfect reference normals. To mimic the distribution of reference normals in the real dataset, we blend linearly between the ground truth and the initial normals of the *frog* dataset. Thus, we obtain the real-world initialization for a blending weight $\alpha = 0.0$ and the ground truth for $\alpha = 1.0$. Figure 6.17a shows the results for several choices of α. First, we note that for almost perfect normals, we obtain almost perfect results on the synthetic images. This safety check indicates that our proposed approach operates correctly in principle. More interesting is the red curve ($\alpha = 0.0$) which can be considered as the upper limit of what we can hope to achieve in reality. The histogram of our real-world results—which contain all other sorts of imperfections—in Figure 6.14a is oriented slightly more to the right, but not significantly worse.

Another aspect that we can study independently using synthetic renderings is image noise. We quantize the high dynamic range images into 255 discrete values to simulate a consumer camera—not counting the effects of a response curve—and add Gaussian noise. The reconstructions for varying standard deviations σ are plotted in Figure 6.17b based on ground truth reference normals. Even for high noise levels, the results are much better than on real images. This indicates that image noise is not the main problem with these datasets.



(a) disturbed input normals

(b) disturbed image intensities

Figure 6.17: Results on synthetic images. a) Reconstructions based on input normals of different quality. High α corresponds to nearly perfect data. $\alpha = 0.0$ mimics the real-world normals for the *frog*. b) Reconstructions with ground truth normals as reference, but input images subject to Gaussian noise with standard deviation σ.

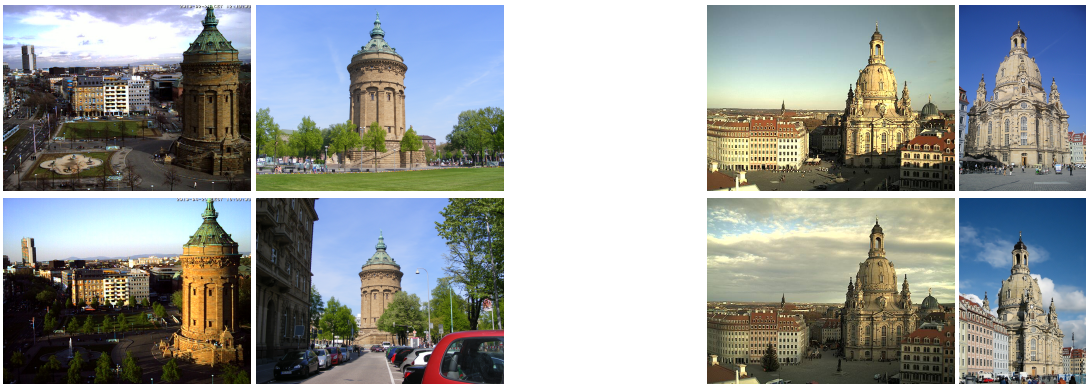


Figure 6.18: Outdoor data. *Left:* Input data for the *tower* dataset. Two images from a static webcam [Webd] (left) and two images taken casually without special capturing setup for the multi-view reconstruction (right). *Right:* Downloaded images for the *church* dataset (left: webcam data [Webb], right: community photos from Pixabay users Simon and clline).

6.5.3 Outdoor Webcam Datasets

We discuss the performance on outdoor scenes using two large buildings with non-planar surfaces and interesting details: the *church* (about 90 m tall) and the *tower* (about 60 m). For each dataset, we retrieve images of a public webcam every 20 min over the course of three months. The webcam images have a resolution of 640×480 pixel, which is typical for this kind of data. In contrast to Chapter 5, we do not apply any kind of calibration since this was one of the motivating scenarios for this chapter. Another difference is that we are not dependent on a specific sky model, *e.g.* clear sky with a bright Sun, and thus can use also images with cloudy skies—as long as there is enough variation overall. We manually select a suitable subset for photometric reconstruction, *i.e.* images taken between 9 am and 7 pm on different days. Some examples are shown in Figure 6.18.

For the multi-view stereo reconstruction of the *tower*, we captured 324 images

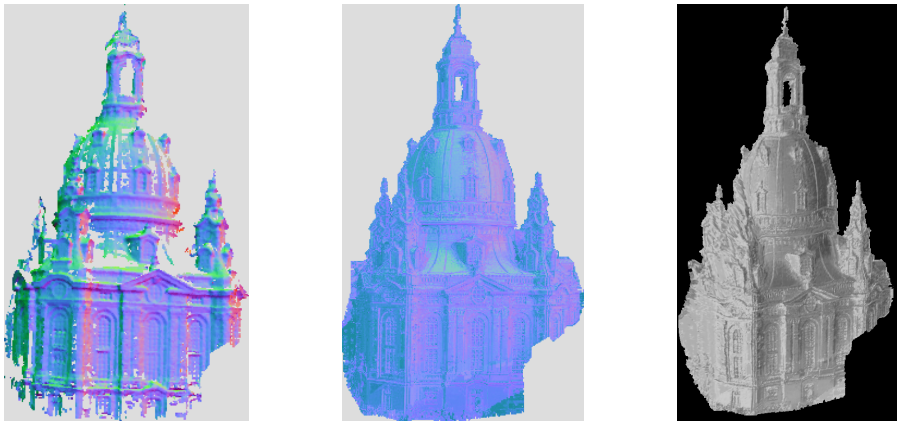


Figure 6.19: Results for the *church* dataset. *Left:* Initial normals. *Middle:* Reconstructed normal map. *Right:* Integrated normal map rendered from a novel view.

with a consumer camera. Since the reconstruction technique [Goesele07] even handles images from community photo collections, we downloaded 2000 images of the *church* from the photo platform Flickr. This gives a dataset which is not only uncontrolled but entirely based on Internet data. As shown in Table 6.1, we obtain a reference that is more complete than in our controlled experiments, but at a much reduced resolution.

We do not have ground truth available for these large-scale outdoor datasets, but provide qualitative results. Figure 6.19 shows the fairly complete reference geometry of the *church*, which is a good basis for reconstruction. The averaging softens the extremes in the final normal map as already observed on the controlled data. We notice the limits of spatial resolution, *e.g.* at the parapet at the base of the dome. There, we observe alternatingly left and right facing normals at frequencies of about one pixel which make the individual pillars almost indiscernible. Larger areas, such as the dome itself, are filled-in plausibly. Artifacts can be seen due to cast shadows on the object that violate the assumptions about illumination and are not modeled by our approach, *e.g.* at the lower right corner.

As a roughly cylindrical object, the *tower* in Figure 6.20 is well-suited for reconstruction. Theoretically, if we parametrize its surface by height $h \in [0, H]$ and angle φ, a normal at (h, φ) can be reconstructed quite accurately if some normal on the line $([0, H], \varphi)$ is contained in the reference geometry. This works so well that the bottom of the tower can be recovered up to fine details such as individual stones. Even though the roof has a different albedo and only sparse coverage of normal directions pointing to the right, we are able to reconstruct it convincingly.

6.6 Discussion

At the beginning of this chapter, we formulated the motivating question whether it is possible to perform photometric shape reconstruction from Internet images without relying on calibrated lights or cameras. The presented approach shows that the answer is indeed positive. In addition, the evaluation gives a more detailed answer on what is possible and what is not. It turns out that different matching scores behave very

Figure 6.20: Results for the *tower* dataset. *Left:* Initial normals. *Middle:* Reconstructed normal map. *Right:* Integrated normal map rendered from a novel view.

similarly and that the results depend very much on the quality of the input geometry.

We conclude from our experiments on uncontrolled data that the fusion of a multi-view stereo initialization with a photometric method works in non-shadowed regions. We expect, however, a fully calibrated approach to yield better results. Managing without a true example object makes several smoothing steps necessary that, in the worst case, could lead to a plane defined by the average normal. Because of this, the performance on Internet data is somewhat disappointing. Especially the normals of the *church* seem to be too "flat". If we consider the challenging capture conditions, however, it is still remarkable to obtain results of this quality.

One of the obvious limitations of the presented approach is that it can only recover normals that are present in the reference geometry—or a convex combination of those. For lower thresholds τ_{conf}, this is less of a problem, but on the other hand leads to more inaccurate normals in the reference which will also be propagated to the reconstruction. For our experiments, we had to find an acceptable trade-off between the number of normal directions present and their quality. A possible extension could be to use a conservative threshold τ_{conf} to obtain a first reference and then add additional normals not only based on their confidence, but also on how they would change the distribution of normal directions.

More generally, problems can arise whenever the partial reference geometry is not representative for the whole scene in terms of normals or reflectance properties. Consider the case where $\mathbf{Q}$ contains only surface points of a certain reflectance, but other parts of the surface show a second reflectance that differs by more than a scalar factor. Then, with the presented matching scores, it would be difficult to obtain a correct match in these parts even for points $p \in \mathbf{P}, q \in \mathbf{Q}$ with the same normal. This is a common problem in many data-driven models and not limited to image-based reconstructions, *e.g.* [deAguiar10]. We cannot expect to make an accurate prediction about a certain pattern or aspect that was never present in the reference or training data. These effects can be reduced in part—but of course not truly overcome—by more sophisticated interpolation schemes.

In abstract terms, we are concerned with a mapping $f : \mathcal{A}_{\mathbf{Q}} \to \mathcal{S}$ from appearance profiles to normals. Our reconstruction goal can then be formulated as extending

f into a mapping $f : \mathcal{A}_{\mathbf{P}\cup\mathbf{Q}} \to \mathcal{S}$ where $\mathcal{A}_{\mathbf{Q}} = \{A_q | q \in \mathbf{Q}\}$ is the set of appearance profiles corresponding to the reference geometry and $\mathcal{A}_{\mathbf{P}\cup\mathbf{Q}}$ is the set of all appearance profiles. Seen from such a general point of view, we can imagine connections to other fields and techniques such as scattered data interpolation, radial basis functions, or kriging. We can also formulate this as a machine learning task with the training set $\mathbf{Q}$ and a test set $\mathbf{P}$ that we would like to annotate with normal information. Applying techniques from these areas to extend the function f might be a direction for future work.

Probably the most obvious connection is with regression analysis. We could formulate an explicit model for f inspired by the approach we used in Chapter 5. From the samples $\mathcal{A}_{\mathbf{Q}} \times \mathcal{S}$, we would have to estimate the illumination L_i in each image and a set of basis materials. The extension of f could then be realized by rendering spheres with these basis materials under the estimated lighting conditions and using mixtures of those as synthetic reference objects during matching. However, we have already observed in Chapter 5 that it is very challenging to robustly estimate such a decomposition. Furthermore, this kind of inverse rendering approaches depend on several design choices and parameters, *e.g.*, the representation of reflectance. In contrast, orientation consistency does not rely on any explicit formulation. This makes it applicable more widely on the one hand, but on the other hand can be too general if a lot of data is missing. In these cases, a stronger model that incorporates more knowledge about the image creation process is appropriate—usually at the cost of more extensive calibration.

No matter which technique is applied, it is also important to consider that the reference geometry only provides a set of noisy samples of the function f. This challenge is not taken into account by traditional example-based approaches, but appears as one of the limiting factors in our evaluation. The intensities in an appearance profile are influenced by global shading effects, and the corresponding normals originate from a noisy multi-view stereo reconstruction. Our proposed averaging of best matching results helps to reduce these effects, but also leads to "flattened" geometry and does not yet take the distribution of candidate normals into account. Developing a noise model and a more principled handling of these inaccuracies during reconstruction is another promising route for further research. Related to this are questions such as "How does the sampling of $\mathcal{A} \times \mathcal{S}$ provided by the reference geometry influence the reconstruction quality?" or "Can we automatically detect outliers such as shadows or interreflections?". To answer these, a better understanding of the space of appearance profiles, their associated normals, and their distribution is needed. A starting point is already provided by Sato *et al.* [Sato07], who apply a dimension reduction technique to the space of observation vectors to reveal an inherent structure.

Future Work: Apart from those already mentioned, some topics for future work come to mind. First, we currently select images from the webcam sequences by hand. Developing a selection scheme, inspired by the image filtering presented in Chapter 5, would enable us to handle larger amounts of data automatically. Second, if we interpret the presented method as a means to fill wholes in the reference geometry, connections to purely geometric hole filling algorithms, *e.g.* [Bendels05], arise. A hybrid approach would have to balance the geometric reasoning, *i.e.* the shape descriptor matching, against the photometric information, *i.e.* appearance profile matching.

Third, it would be interesting to push the development towards Internet reconstructions by acquiring ground truth for a large-scale outdoor dataset. A quantitative evaluation in such a setting would permit a better understanding of the challenges that have yet to be addressed. Finally, we presented a working system comprised of several components, such as initial geometry reconstruction, matching, normal transfer, *etc*. For most of these, we chose a straightforward solution. On the one hand, this makes the approach easier to implement than the pipeline in Chapter 5. On the other, each of these steps can easily be replaced and augmented, for example, by a shadow detection step or by introducing a sophisticated confidence measure during integration.

Chapter 7

Multi-View Photometric Stereo by Example

Both approaches we have discussed in the previous chapters use static webcams as the source of input images. This avoids the problem of unknown dense pixel correspondences between images. For the same reason, most photometric stereo techniques assume a fixed camera. This restriction is one of the factors that prevents their application on general Internet photo collections that do not contain webcam sequences. Even in the lab, where it is easier to keep the camera fixed, reconstructing only one side of an object is still an undesired limitation. For example, a slanted surface in one camera might be observed almost fronto-parallel in another and thus introduces less distortion and yields more details in the reconstruction. We therefore present an approach that works for images from different view points. It still exploits photometric properties, which sets it apart from multi-view stereo techniques that rely on color constancy.

The multi-view case is a lot more challenging than regular photometric stereo, even under controlled conditions, because of the additional degrees of freedom. We consider a lab setting for this chapter to simplify at least some aspects of the problem, mentioned in Section 1.1.2. Thus, we will not achieve our ultimate goal of reconstructing an object from uncontrolled Internet images, but we keep the specific challenges of this data in mind and let them influence our design decisions. In particular, we would like to keep the calibration effort low in terms of lighting and radiometric response curve. Additionally, we would like to allow complex non-Lambertian reflectance and arbitrary camera placement. Our focus is on textureless surfaces because classical (multi-view) stereo methods have problems on these. Reconstructing accurate geometry for such objects is still a very challenging task under unknown lighting conditions if no special setups such as ring lights or calibration steps are employed.

In recent years, first multi-view photometric stereo techniques have been presented as discussed in Section 3.6. Many approaches begin by reconstructing a proxy geometry which provides correspondences for photometric stereo. Common choices to obtain the proxy geometry are depth maps from structured light (Zhang *et al.* [Zhang12]), a piecewise-planar initialization constructed from tracked feature points (Lim *et al.* [Lim05]), multi-view stereo reconstructions (Park *et al.* [Park13]), simple primitive meshes (Yoshiyasu and Yamazaki [Yoshiyasu11]), and the visual hull computed from silhouettes (Hernandez *et al.* [Hernandez08]). However, in general,

feature extraction and stereo reconstructions fail on textureless surfaces. The visual hull only provides an adequate initialization if the object is observed from considerably varying angles. None of these approaches uses photometric cues for depth estimation. Instead, they use it only to recover normals which then affect depth indirectly. In contrast, we are motivated by the question whether depth can be reconstructed from the information encoded in the changing image intensities due to variations in lighting.

As mentioned, our goal is to recover the surface of objects with possibly non-Lambertian BRDFs. Most multi-view approaches treat non-Lambertian reflectance as outlier. This works reasonably well, but can only go so far and usually relies on a large amount of images. We handle complex BRDFs more explicitly but without restricting our approach to a certain parametric model. To achieve this, we place a reference object with known geometry and the same reflectance properties as the target in the scene. Although such an example object limits the practical applicability, it makes our technique independent of a specific BRDF model. Furthermore, it makes a calibration of the camera response unnecessary which is required by many other photometric stereo techniques. Similar to the example-based approach in Chapter 6, we then match per-pixel appearance profiles under different point light illumination—and now also varying view point—between the target and reference object. While the matching error turns out to be not very discriminative for reconstruction of depth, we show that normals can be recovered very reliably in the vicinity of the true surface. That insight leads to an energy formulation which incorporates the recovered normals as soft constraints.

This approach eliminates several restrictions of the voxel coloring-based work by Treuille *et al.* [Treuille04]. They also match appearance profiles as introduced by Hertzmann and Seitz [Hertzmann03] and use the error as consistency measure in a voxel coloring framework. Their formulation has, however, several drawbacks: First, it poses restrictions on camera placement to ensure that occluded voxels are processed in the correct order. We allow arbitrary (distant) camera placements and rely solely on generic outlier removal to handle occlusions. Second, their final scene representation is a voxel grid. The reconstruction cannot be easily transformed into a surface while taking the normals into account, *e.g.* extracting a surface using marching cubes still yields "blocky" geometry. In contrast, we use a local surface representation with continuous depth values and obtain smooth results that contain all details. Finally and most importantly, their approach does not use the reliable normal information during depth recovery, which makes it prone to errors in the reconstructed geometry. Our approach differs from [Treuille04] in scene representation (voxels vs. multiple depth maps), visibility model (geometric vs. outlier-based), and the reconstruction algorithm (voxel coloring vs. per-view non-linear optimization).

7.1 Problem Statement

Our goal is to recover the surface of a textureless object solely from a set of images $\{I_1, \ldots, I_M\}$ under varying illumination and from different viewpoints. We also want to keep the capture procedure simple and straightforward. In practice, this means that we avoid any calibration of light sources or camera response curves. If we also allow for non-diffuse surfaces, none of the existing techniques can be applied. We base our approach on orientation consistency as a depth cue, which brings many of

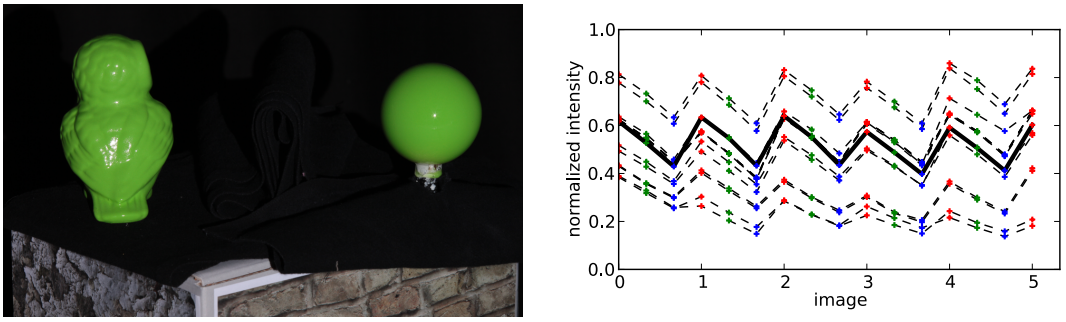


Figure 7.1: *Left*: Target object and a reference sphere with same reflectance. The high frequency pattern at the bottom is used to estimate camera pose. *Right*: Some samples from the database of reference profiles (dashed) and a candidate profile (solid).

the desired properties, and thus place a reference object with known geometry in the scene. Figure 7.1 (left) shows an example.

The surfaces in a scene can be described as a global model, *e.g.* voxel grids, triangle meshes, *etc.*, or local representations such as per-view depth maps. We choose the latter option which makes parallel processing of individual views trivial. Let $I = I_1$ denote a master image and r the ray corresponding to pixel p. We assume that the camera projection matrices $\{P_1, \ldots, P_M\}$ are known. For a candidate depth z, we project the corresponding 3D position $z \cdot r$ into all M images to obtain intensities $I_i(P_i(zr)), i \in \{1, \ldots, M\}$, in each of the three color channels. The concatenation of the $3M$ values into a vector $A(zr)$ forms an *appearance profile*. In contrast to the definition in Section 6.3.1, we do not group the individual color channels.

As a reference object, we use a sphere with the same reflectance properties as the target object. For now, let us assume that position and radius of the sphere are known. For each pixel in I that is covered by the sphere, we project the corresponding sphere point into all images and form a reference appearance profile B. This gives a database of profiles with attached normals $\tilde{n}$ computed from the sphere. Some of these reference profiles B together with a candidate profile A are visualized in Figure 7.1.

We assume a distant but otherwise unknown point light source. Shadows and interreflections are handled as outliers during matching without explicit treatment.

7.2 Approach

7.2.1 Matching

Assuming an orthographic camera, the intensity of a surface point zr with normal n is given by Equation (3.35) as

$$I_i(P_i(zr)) = f_i \left(\int L_{s,i}(\omega) \rho(\omega, v_i, n) \langle n,\ \omega \rangle d\omega \right) \tag{7.1}$$

where f_i is the camera response, ρ the BRDF, $L_{s,i}$ describes the point light source, and v_i the camera viewing direction. Both, light and camera position, change from image to image as indicated by the index i. *Orientation consistency*, as discussed in Section 3.5, states that the intensity is the same for a point with the same normal on the reference object.

This means that we can find a matching profile B in our database for any $A(zr)$ that originates from the true surface. For a false depth candidate z, it is unlikely to find a good match because each view actually observes a different point on the surface (see Figure 7.3 for an example). We denote the intensity residuals $e_i = A_i - B_i$ and omit the color channel indexing.

In Section 6.3.1, we discussed several ways to define a matching error on appearance profiles. The evaluation did, however, not result in a clear vote for one over the others. In the multi-view setting, we expect an increase of outliers in the target profiles due to occlusions, which do not occur for a fixed camera. Treuille *et al.* [Treuille04] use the normalized L_2 distance as a matching error. The contribution of e_i is not considered during matching if the corresponding voxel was actually occluded in image I_i. We do not have occlusion information available for the components of the target profiles A. Instead, we turn off residuals e_i if the corresponding normal to the reference B has been observed at a grazing angle in the i-th view. Furthermore, we only use the K best of the remaining residuals (set as a percentage, typically 60%):

$$E_{\text{match}}(A, B) = \frac{1}{K} \sum_{j=0}^{K} e_{i_j}^2 \tag{7.2}$$

where $e_{i_1} \leq \ldots \leq e_{i_M}$. If K is lower than three, we set $E_{\text{match}}(A, B) = \infty$ because normals cannot be recovered unambiguously with fewer lighting variations.

7.2.2 Energy Formulation

Along a ray r, the best matching error at position zr

$$E_M(r, z) = \min_B E_{\text{match}}(A(zr), B) \tag{7.3}$$

gives an indication whether we are on the true surface or not. Unfortunately, the matching error is not very discriminative as shown in Figure 7.2a. We do not observe a clear minimum but rather depth values with a wide basin of low error. Accordingly, choosing the depth with smallest matching error leads to a very inaccurate and noisy depth map like in Figure 7.2b. The standard way to deal with noise and unreliable estimates, *e.g.* in stereo, is to employ regularization that favors smooth surfaces. We have the advantage of additional information in the form of normals associated with the best match from the database. To exploit these, we formulate an energy that is defined on both a depth map Z and a normal map N. This can be interpreted as attaching a small oriented plane $(Z(p), N(p))$ to each ray, see Figure 7.3 (left).

The key finding in our setting is that exactly the same reasons that make depth estimation hard, make normal estimation easy. Figure 7.3 (right) illustrates this insight for three different points along the same ray. In Figure 7.3a, all cameras observe the same point on the true surface. The matching error will be low and the normal $\tilde{n}$ associated to the match is the correct surface orientation n. If we move slightly away from the surface as shown in Figure 7.3b, each camera actually observes a different surface point but with normals that are still close to the true one. Accordingly, the intensity profile will be very similar to the previous one. Thus, the matching error is again low which makes accurate depth estimation so difficult, but the associated normal is close to n. This reasoning breaks down if the point is really far away from

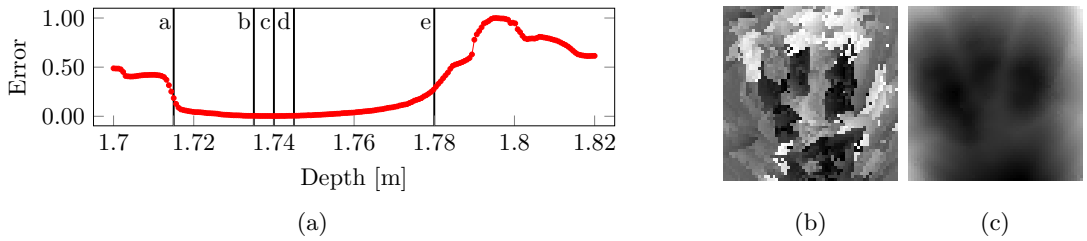


Figure 7.2: a) The error of best matching reference profiles along a ray from the camera has a wide basin with very similar error scores. The vertical lines correspond to the depth values in Figure 7.4. b) Selecting a discrete minimum along each ray leads to noisy depth maps. c) Incorporating normals gives a smooth result while keeping the structure intact.

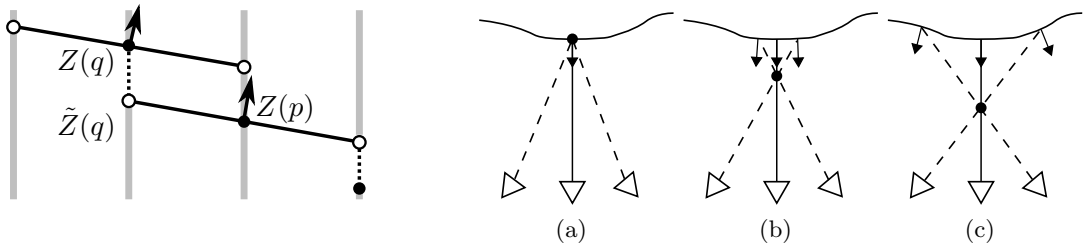


Figure 7.3: *Left*: Each ray has a little plane attached. The estimated depth of neighboring pixels should be close to the intersections of their rays with the plane. *Right*: Projections at different depth along a ray. (a) All cameras observe the same point. The matching error is zero. (b) Cameras observe different points, but with similar normals. The matching error is still low. (c) Cameras observe points with significantly different normals. The matching error is high.

the surface as in Figure 7.3c. All cameras observe surface points with very different normals, and the normal associated with the best match will not be close to any of them. In this case, the matching error itself is high.

Figure 7.4 shows this effect on real data. For a ray indicated by the green dot at pixel p on the target object, the best matching normals are visualized for five depth values corresponding to the plot in Figure 7.2a. We observe that the normals are almost constant in the region of low error. To exploit this finding, we focus our optimization on the normals and use the matching error only as a weak constraint.

Based on these considerations, we propose the following energy formulation

$$E(Z, N) = E_M(Z) + \alpha E_{\text{copy}}(Z, N) + \beta E_{\text{coupling}}(Z, N). \tag{7.4}$$

The first term is the sum of matching errors over all rays for the current depth estimates, which involves matching against the intensity database for a single evaluation of $E_M(r, z)$:

$$E_M(Z) = \sum_r E_M(r, z)^2. \tag{7.5}$$

The second term effectively copies the normal $\tilde{n}$ associated to the best matching

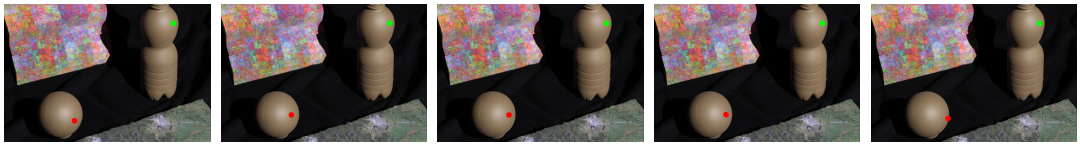


Figure 7.4: Along the ray going from the camera through the pixel marked in green, the normal corresponding to the best matching reference profile is visualized (red) for increasing depth. Images from left to right correspond to depth a-e in Figure 7.2. Close to the surface (depth b,c,d), normals are very stable and similar to the true one.

reference profile to the estimated normal map but also allows for deviations:

$$E_{\text{copy}}(Z, N) = \sum_r \|n - \tilde{n}\|^2. \tag{7.6}$$

The best matching $\tilde{n}$ also depends on the depth z, which we omitted here for clarity. Internally, we parametrize the normals in angular coordinates to ensure unit norm.

The third term couples depth and normals. We assume that the surface is locally planar at a pixel p, but not necessarily fronto-parallel. We look at a neighboring pixel $q \in \mathcal{N}(p)$ and intersect its ray $r(q)$ with the plane defined by $(Z(p), N(p))$:

$$\tilde{Z}(q) = Z(p)\frac{\langle r(p),\ N(p)\rangle}{\langle r(q),\ N(p)\rangle} =: Z(p)\frac{s(p)}{s(q)}. \tag{7.7}$$

The intersection point $\tilde{Z}(q)r(q)$ should then be close to the current estimate $Z(q)r(q)$ as shown in Figure 7.3, and we obtain the following coupling term

$$E_{\text{coupling}}(Z, N) = \sum_p \sum_{q \in \mathcal{N}(p)} E_{\text{coupling}}(p, q), \tag{7.8}$$

$$E_{\text{coupling}}(p, q) = \left(Z(p)s(p)/s(q) - Z(q)\right)^2. \tag{7.9}$$

The coupling energy is similar in spirit to the orientation constraint used by Nehab *et al.* [Nehab05], who compare normals with local tangents. We allow for varying N at the cost of loosing linearity, but the matching error E_M is non-linear anyway. Furthermore, our formulation avoids an explicit numerical differentiation. Those are often based on central differences for increased stability but can lead to an unwanted decoupling of neighboring pixels.

Approaches that start with a proxy geometry and then obtain the final surface through a single refinement step, as proposed by Joshi and Kriegman [Joshi07] or Park *et al.* [Park13], exploit the additional knowledge about the surface orientation only in this final phase after fundamental decisions on depth have already been made. This can lead to problems if the initialization is inaccurate as in our case. Therefore, we relate depth and normals to the input intensities in one optimization problem and allow all decisions to be made at the same time.

7.3 Implementation

So far, we have looked at the theoretical description of our approach. To apply it in practice and run real experiments, some further remarks are appropriate.

Dataset	Pixels in Mask	Energy after Iteration 0	10	50	Time [min]
Bottle	29k	3263	1129	164	459
Diffuse Owl (front)	48k	7712	2408	562	286
Shiny Owl	13k	12331	274	46	130
Spheres	12k	589	49	47	41

Table 7.1: Computation times and optimization performance for the datasets and views discussed in Section 7.4.2.

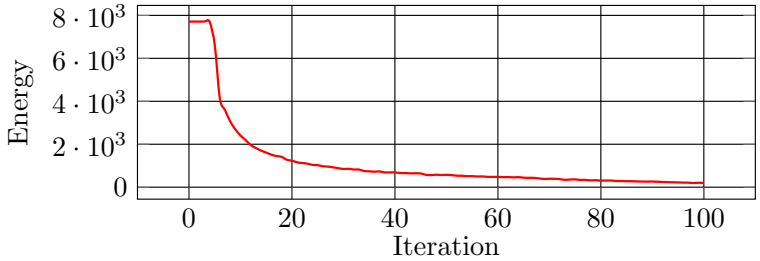


Figure 7.5: Convergence behavior (shown exemplarily for the view in Figure 7.8 (top)). The energy decreases with growing number of iterations. At 50 iterations, it has dropped by more than one order of magnitude.

Optimization: We use the Ceres non-linear optimization package [Agarwal] to minimize the energy in Equation (7.4) with the Levenberg-Marquardt algorithm. Our formulation is, however, non-convex and has many local optima. It is therefore crucial to obtain a sufficiently good initialization. We define a depth range which we sample in discrete steps similar to a plane sweep and evaluate only the term E_M. For each pixel, we use the depth that results in the lowest error and copy the corresponding normal from the reference object. As already mentioned, these estimates are rather noisy in depth. Still, the normals provide a suitable starting condition. Furthermore, we allow the solver to make jumps that temporarily increase the energy if it ultimately leads to a smaller error. This helps to avoid local optima at the cost of increased run time. In our experiments, we stopped the optimization after 50 iterations even if the solver did not fully converge. We found this to be a reasonable trade-off between quality and computation time. Better results are possible at the cost of a greater time budget. Table 7.1 and Figure 7.5 show that the energy is decreased by one to two orders of magnitude with our current setting. In our prototype, we use images of size 1400×930 and 700×465. This is to reduce run time since the main bottleneck lies in the matching of each candidate profile against all reference profiles. Acceleration with spatial data structures is difficult because our matching is not a true metric due to the outlier tolerance.

Assumptions in Practice: In Section 7.1, we made the assumptions that camera parameters and the position of the reference sphere are known. To obtain these parameters, we place a target with a high frequency texture in the scene, see Figure 7.4. We then extract features and apply structure from motion followed by bundle adjust-

(a) (b) (c) (d)

Figure 7.6: Datasets with varying reflectance. *Left to right*: Input images for the *bottle*, *diffuse owl*, *shiny owl*, and *spheres* datasets corresponding to the depth and normal maps discussed in Section 7.4.2. We use the high frequency patterns placed in each scene to estimate camera pose.

ment. The reference sphere is located by fitting conics to the outline of the sphere in the images. Afterwards, the rays through the sphere center are intersected to find its position. This procedure has the additional advantage of providing us with metric scaling information based on the known radius of the sphere. The metric coordinate system then helps to define the depth range during initialization of the optimization.

Preprocessing: Including all possible images in the reconstruction of a given master view not only leads to increased processing cost, but can also reduce robustness. If the parallax between two views is too large, chances are that they actually observe different parts of the surface. We avoid measuring consistency between such views and automatically discard images with a viewing direction that deviates more than 50° from the master view. As another preprocessing step, we manually define a mask for the object in the master view.

Parameter Settings: The weighting factors in Equation (7.4) are chosen according to the range of each subterm. The input intensities and E_M are in $[0, 1]$. E_{copy} is in $[0, 2]$ since we do not enforce front-facing normals. We assume that depth is measured in meters, but the typical deviations between neighboring pixels are only fractions of millimeters. Therefore, we scale E_{reg} to lie in a similar range as E_M and E_{copy}. In summary, we set $\alpha = 1$ and $\beta = 5000$ in all our experiments. For much larger β, the surface moves away from its true position, whereas for much smaller values it remains noisy. Another parameter is the depth range for the initialization. We manually select a range that encloses the object by 10 cm to 15 cm and sample it in 200 steps.

7.4 Evaluation

There are no multi-view datasets available for benchmarking that contain a reference object. Thus, we captured four real-world datasets of objects with varying degrees of specularity. We present a qualitative evaluation by showing the recovered depth map, normal map, and a novel view of the triangulated geometry. We also analyze the results for depth and normals quantitatively.

7.4.1 Experimental Setup

For all experiments, we used a point light source at a distance of 4.5 m to approximate distant illumination. We placed the reference and target objects close together to ensure equal lighting conditions. Figure 7.6 shows some examples of our input images. The *bottle*, *shiny owl*, and *spheres* datasets were captured by moving the camera and light source in each shot. For the *diffuse owl* dataset, we captured views from 360° using a turntable and varying the height of the camera and light. It consists of 39 images, whereas the other datasets contain only about 15 images each. We used a professional-level camera (Canon EOS 5D) except for the *bottle* dataset, which was captured with a consumer camera (Canon EOS 700D). The corresponding lenses have focal length 135 mm and 160 mm (in 35 mm equivalent) and approximate an orthographic camera. All results are computed on non-linear JPEG images. We intentionally did not apply gamma correction since dealing with non-linear intensities is one of the strengths of our technique. The spheres are made of plastic or glass and have diameters of 3.5 cm to 4 cm. We accept imperfections such as a gluing edge or small dents at the benefit of inexpensive and readily available reference objects.

7.4.2 Overall Results

We first present qualitative and quantitative results for each of the datasets before looking at specific aspects in more detail. In order to create a textureless target object, we sprayed a bottle and an example sphere with brown paint such that they have a BRDF with a broad highlight but are not truly diffuse (see Figure 7.6a). The shape of the bottle is rather uniform and can be recovered quite well as shown in Figure 7.7. Even the fine grooves are visible in the normal map and the triangulated depth map. In the concave middle section, some problems occur if the amount of shadowed images exceeds the outlier tolerance of our matching norm. We also acquired a ground truth model for the *bottle* dataset with a structured light scanner and manually registered it with our input images. The bottom of Figure 7.7 shows two planes that cut through the ground truth and our depth map. We observe that the ground truth and our reconstruction are slightly shifted. The deviations are smaller than 4 mm which is at the scale of the alignment error, given that the camera was 2 m distant.

The *diffuse owl* is a 12 cm tall porcelain figurine which we spray painted with a diffuse green color to create a homogenous reflectance, see Figure 7.6b. The depth of the left wing in the top row of Figure 7.8 is estimated too small, probably because it is occluded in most of the cameras to the right. Overall, the reconstructed normal map and the rendering of the triangulated depth show fine details for both views and only some artifacts at depth discontinuities, such as the bottom right corner of the eye. After we captured the *diffuse owl* dataset, we applied a transparent varnish to the figurine which makes it appear glossy as shown in Figure 7.6c. This novel *shiny owl* dataset demonstrates our performance on non-diffuse surfaces. Even small details such as the feathers on the tail are clearly recognizable in Figure 7.9.

As a second example of shiny reflectance, we use two transparent Christmas balls lacquered from the inside with acrylic paint, see Figure 7.6d. We use the left one as reference object and reconstruct the one on the right. This way, we can quantitatively compare the reconstructed normals in Figure 7.10 against those of an ideal sphere whose position we obtain as described for the reference sphere. Although the target is

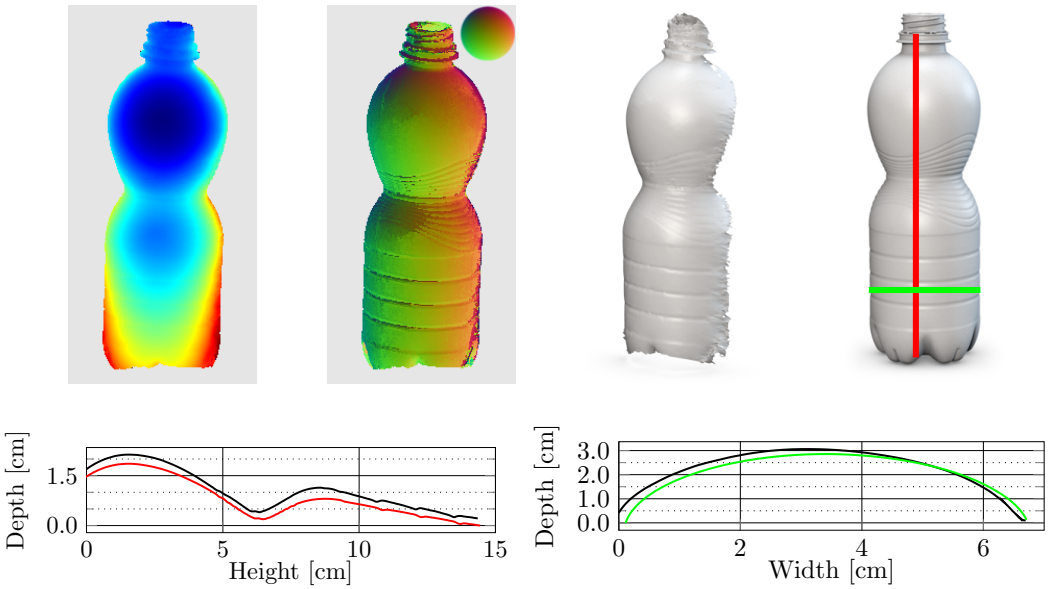


Figure 7.7: Results for the *bottle* dataset. *Top, left to right*: Colored depth map from blue (near) to red (far), the normal map, a rendering of our triangulated geometry from a novel view, and the ground truth acquired from structured light scanning with profile lines from left to right (green) and top to bottom (red). *Bottom*: The vertical (left) and horizontal (right) cuts through the ground truth (red and green) and our depth map (black) show a deviation of less than 4 mm.

not perfectly round and its reflectance does not completely match the reference due to varying thickness of the dye coating, the angular deviation is low. The histogram in Figure 7.10 is peaked at 5° with most of the larger errors—besides at the boundaries—originating from the sphere center where the over-exposed highlight was observed most often. A greater variation of light directions should fix this.

7.4.3 Optimization Performance

So far, we have looked at final results of the optimization. We now also consider other aspects of our pipeline, *e.g.* the impact of the optimization. Figure 7.11 contains an additional view of the *shiny owl*. It shows that, while the initialization already provides good normals in many places, it has lots of incorrect depth estimates. This is in accordance to our discussion in Section 7.2.2. The optimization result is a clear improvement both in depth and normals. We stop the non-linear solver after 50 iterations and have shown in Figure 7.5 that this already leads to a drastic decrease of the overall error. We now study how this relates to the depth and normals over several iterations. Figure 7.12 shows the changes made by the optimization on the *diffuse owl* dataset. The difference between 50 and 100 iterations is small and noticeable only in "extreme" regions, *e.g.* around the feet. Neither 50 nor 100 iterations are able to find the correct normals at the lower right and upper left corner of the eye. Running more iterations will not solve this problem which is probably caused by incorrect matches that nevertheless produce low error.

These results for the *diffuse owl* indicate that the solver is either stuck in a local minimum or that our energy does not model reality in the way we would expect.

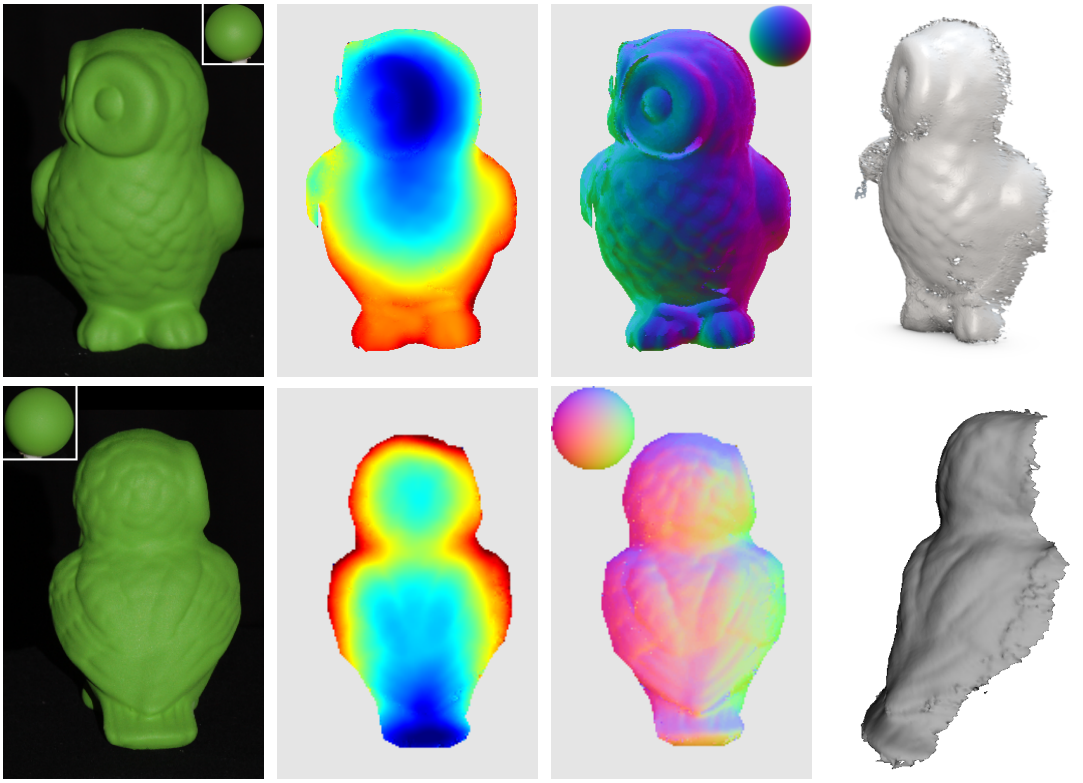


Figure 7.8: Results for the *diffuse owl* dataset from two different camera positions (top: computed from 1404×936 pixel images, bottom: computed from down-scaled images). *Left to right*: Cropped input image with the reference sphere as inset, colored depth map from blue (near) to red (far), the normal map, and a rendering of our triangulated geometry from a novel view.

To ensure that the energy does in fact have a minimum close to the true surface, we initialize the optimization with ground truth data. For an additional view of the *spheres* dataset, we render the depth and normals of an ideal sphere at the position of the target and use this information to initialize the solver. The top row of Figure 7.13 demonstrates that, as expected, the result after 50 iterations moves away only slightly from the start configuration. The bottom row shows that comparable results are obtained with a realistic initialization.

We encountered some issues that seem to be related to the scale of the input images. The results in the top row of Figure 7.8 are computed on 1404×936 pixel images and perform similarly on the down-scaled version (702×468 pixel). The view in the bottom row is shown for the smaller version and also works very well. On the other hand, the same view is a failure case if computed on 1404×936 images. Figure 7.14 shows the optimized depth and normal maps together with two renderings that show a warped geometry. The images in the bottom row demonstrate that the initialization is very similar in the large and the down-scaled version. The solver runs into an incorrect local optimum, probably because the erroneous regions still have homogenous and locally consistent normals. With fewer unknowns, the coupling error at the boundary between correct and incorrect depth becomes more important and helps to guide the optimization. Experimenting with stronger coupling constraints or

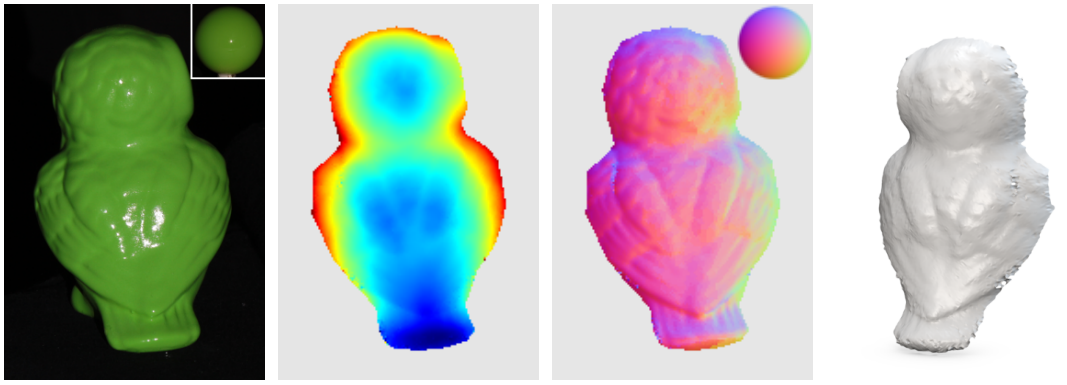


Figure 7.9: Results for the *shiny owl* dataset. *Left to right*: Cropped input image with the reference sphere as inset, colored depth map, the normal map, and a rendering of our triangulated geometry from a novel view. Even for shiny surfaces, fine details can be recovered.

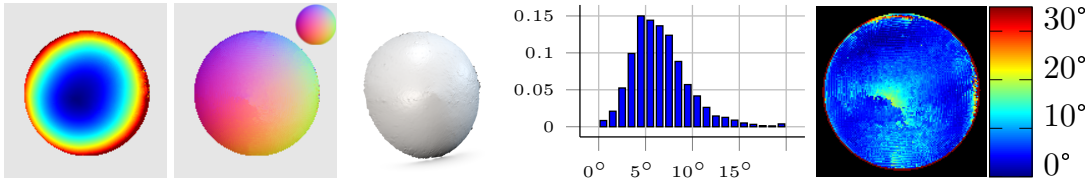


Figure 7.10: Results for the *spheres* dataset. (a,b,c) The reconstructed depth map, the normal map, and a rendering of our triangulated geometry from a novel view. (d,e) The normalized histogram of angular errors below 20°, and the spatial distribution of angular errors. We compute angular errors by comparing the reconstructed normals to those of an ideal sphere.

incorporating an additional smoothness term that operates on larger structures than single pixel neighborhoods are interesting directions for future work.

7.4.4 Consistency of Local Reconstructions

We reconstruct per-view depth and normal maps. This is sufficient for many tasks, *e.g.* image-based rendering [Goesele10a]. If a global model of the object is desired, these local surface representations have to be merged. Since this is an often occurring problem, many techniques exist to fuse multiple depth maps into a surface, *e.g.* [Curless96, Kazhdan06, Fuhrmann11]. Of course, a certain amount of consistency between the local representations is needed for successful merging. In our case, achieving consistent depth maps might pose a particular challenge because we rely a lot on the information present in the normals. Integrating normal maps can result in globally deformed surfaces if not sufficiently constrained by depth information, see [Klette96] or the discussion in Section 6.4.

The red and blue vertices in Figure 7.15 (left) demonstrate that our reconstructions are consistent for views that are close to each other. The green vertices belong to a view that is rotated more heavily and thus leads to larger deviations. Still, they are in the order of only a few mm. For all reconstructions, we notice that the discrepancy is most pronounced at the boundaries. In these regions, even small alignment errors between

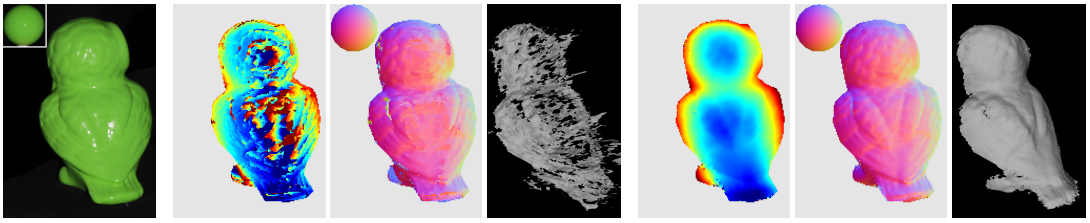


Figure 7.11: Improvement through optimization. *Left*: The input image corresponding to the view used in this figure (differing from Figure 7.9). *Middle*: The color coded initial depth map, the initial normals, and a rendering to demonstrate the inaccuracy of the initialization. *Right*: The final depth map, normal map, and a rendering from a novel view.

cameras can lead to wrong intensities being observed during matching. Cropping away the boundary region in each of the individual depth maps before merging might be an option. Figure 7.15 (right) shows a global mesh that we fused from 17 views of the *diffuse owl* without any editing. All depth and normal maps were projected to oriented 3D points and then processed using Poisson Surface Reconstruction [Kazhdan06]. Some details are lost compared to the per-view reconstructions. More images in the dataset and a merging technique specifically tailored to depth maps will give better results but are left for future work.

7.4.5 Comparison to Voxel Coloring

Matching appearance profiles in a multi-view setting has also been studied by Treuille *et al.* [Treuille04]. Unfortunately, that work does not contain a quantitative evaluation that we could compare against. We therefore reimplemented their technique and show the results in Figure 7.16. The *diffuse owl* dataset contains views from all directions, and voxel coloring produces a reasonable reconstruction. It is, however, discretized and detail information encoded in the normals is only accessible for rendering. In contrast, our energy formulation is continuous in depth. It thus leads to a fundamentally different optimization problem. We provide a quantitative comparison with our reconstruction for the *bottle* where ground truth is available. This dataset contains only 14 cameras that observe the object mostly from the front. As a result, the voxel reconstruction is not able to recover the true shape because the matching error is not very discriminative. In contrast, our approach enforces consistency of reconstructed normals and depth, which provides a clear advantage.

7.4.6 Different BRDF on Reference and Target

Orientation consistency relies on the reference and target to have the same reflectance. An interesting question is how large the impact of this restriction is in practice. We captured an additional dataset that contains the brown bottle (*bottle2*) and the white Spectralon sphere introduced in Section 4.4. The latter is almost perfectly Lambertian. The brown paint, on the other hand, has a specular component, as shown in Figure 7.17, and therefore reflects light quite differently. We manually adjusted the albedo in the appearance profiles of the bottle to approximate a white color. This does not change the reflectance behavior and does in particular not change the (occurrence of) the specular highlight. Adjusting the albedo is only strictly valid if we drop the

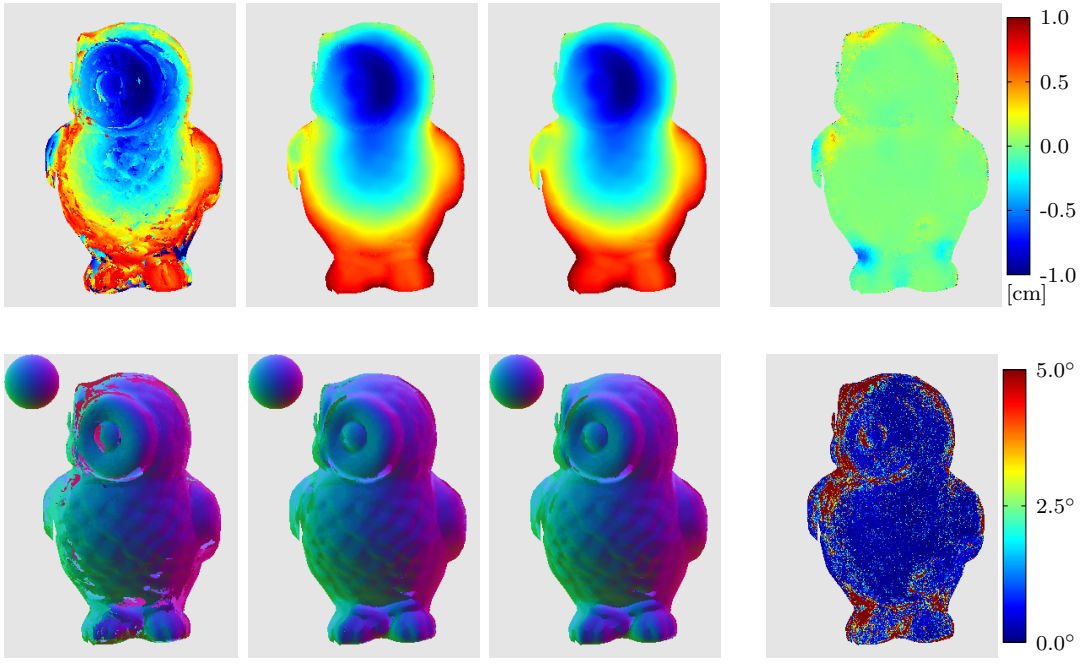


Figure 7.12: Changes in depth (*top*) and normals (*bottom*) during optimization of the *diffuse owl*. *Left to right*: The initialization, the result after 50 iterations, the status after 100 iterations, and the difference between 50 and 100 iterations.

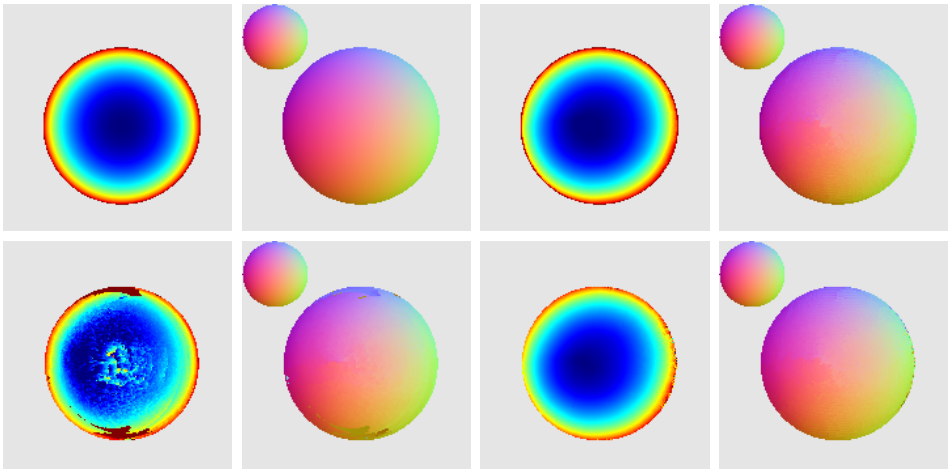


Figure 7.13: Influence of initialization. *Top*: Initialization with an ideal sphere. Left to right: Initial depth and normal map, optimized depth and normal map. *Bottom*: Initialization through discrete evaluations along each ray. Left to right: Initial depth and normal map, optimized depth and normal map.

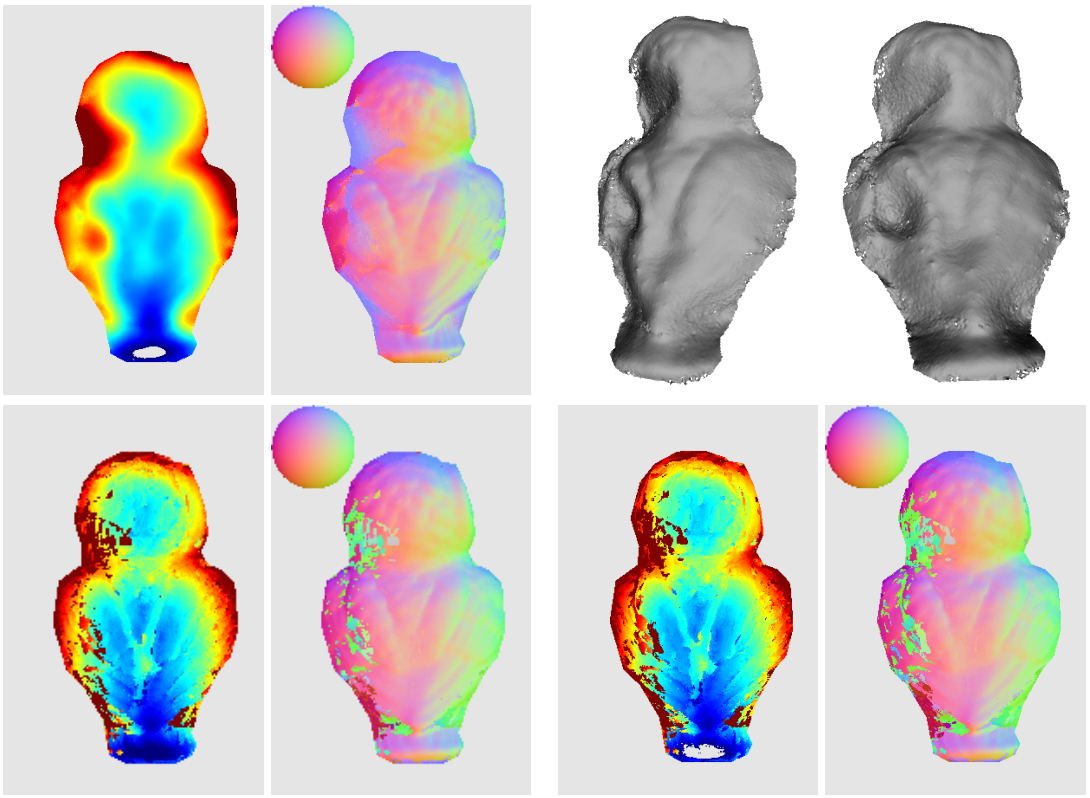


Figure 7.14: Reconstructing the same view of the *diffuse owl* as in Figure 7.8, but from larger images, fails. *Top*: Results computed from 1404×936 pixel images. Left to right: The reconstructed depth map, the corresponding normals, and renderings of the triangulated geometry from two novel views. *Bottom, left*: The initial depth and normal map used for the reconstruction in Figure 7.8 on images of size 702×468. *Bottom, right*: The initial depth and normal map for 1404×936 pixel images.

assumption of a possibly non-linear camera response, but we found it to work even on the JPEG images we use. Figure 7.18 shows that, for this example, our matching-based approach is able to cope even with differences in BRDF between the target and the reference sphere. The result is comparable to the *bottle* dataset (see Figure 7.7), for which target and reference had the same reflectance.

7.5 Discussion

On the path to more general photometric stereo methods, the approach presented in this chapter removes the very common assumption of a fixed camera view. Other works towards that direction recover geometry with non-photometric techniques, such as stereo, and are rather "multi-view X *plus* photometric stereo" than "multi-view photometric stereo" methods. In contrast, our motivation was to investigate whether shading variation due to illumination changes alone is sufficient for depth reconstruction. We achieved that goal using orientation consistency inspired by Chapter 6.

For scenes that lend themselves well to patch-based comparisons or that have enough view variation to extract a sufficiently good visual hull, it makes of course sense to apply a reconstruction technique that exploits this information. In our case

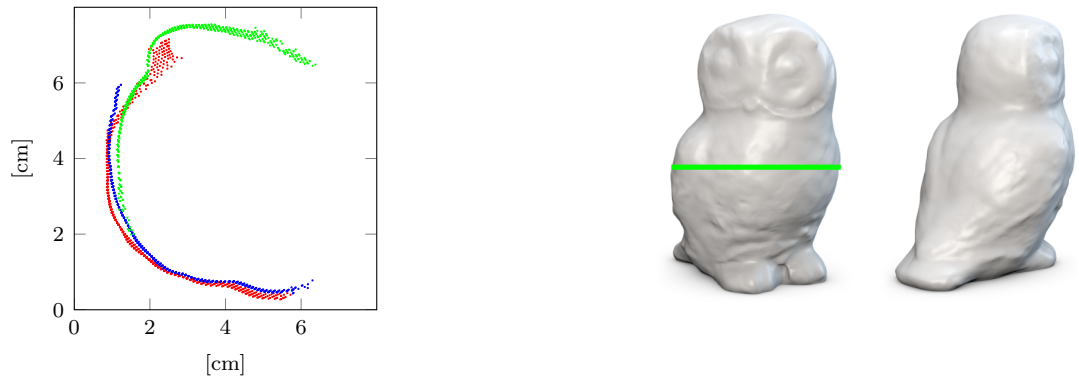


Figure 7.15: Consistency between views. *Left*: Several horizontal slices (corresponding to the green line in the image on the right) of three depth maps in the *diffuse owl* dataset are projected onto a plane. We observe that reconstructions of nearby views (red and blue) are quite consistent. In general, larger deviations occur towards the boundary of individual depth maps. *Right:* Two renderings of a globally consistent model obtained by merging 17 depth maps using Poisson Surface Reconstruction [Kazhdan06].

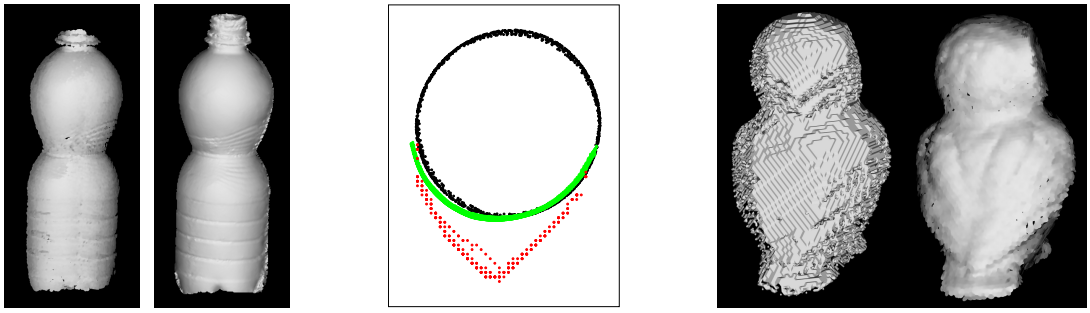


Figure 7.16: Comparison to Treuille *et al.* [Treuille04]. (a) The voxel-based reconstruction of the *bottle* rendered using point splatting. (b) Our reconstruction shown from the same view. (c) Geometry comparison: several horizontal slices through the *bottle* reconstructed with our approach (green), Treuille *et al.* [Treuille04] (red), and structured light (black) are plotted on top of each other. (d) The marching cubes reconstruction of the volume by Treuille *et al.* is blocky as shown for the *diffuse owl* dataset (left). The attached normals do not contribute to the geometry and can only be used for shading (right).

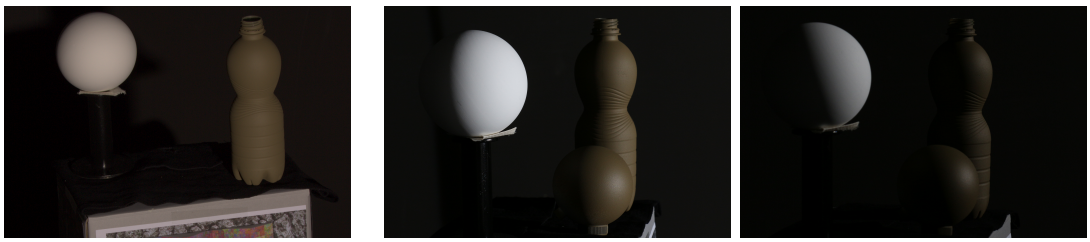


Figure 7.17: Different BRDFs. *Left*: One of the input images for the *bottle2* dataset containing a diffuse sphere and the brown bottle. *Middle and right*: Two images that visualize the difference in specularity. The brown sphere (not used for reconstruction) and *bottle* show a specular highlight around the halfway vector between light and viewing direction. The white Spectralon sphere is purely diffuse.

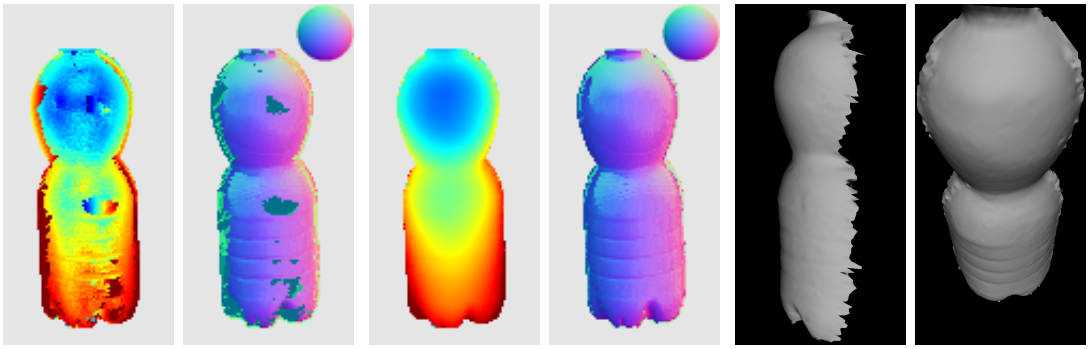


Figure 7.18: Matching with unequal BRDFs. *Left*: Initial normals and depth map for the *bottle2* dataset. *Middle*: Normals and depth map after 40 iterations. *Right*: Two renderings of the triangulated depth map shown from novel camera positions.

of textureless surfaces, we were not able to obtain even an approximate geometry from the multi-view stereo technique by Goesele *et al*. [Goesele07]. Similarly, the visual hull for the *bottle*, computed from less than 15 images which all observe the object from the front, was too far from the surface to act as a suitable proxy. We have shown that using photometric cues for depth estimation is a viable option for these kinds of scenes.

Our algorithm has several desirable properties such as avoiding radiometric calibration but is unfortunately not well suited for Internet images. Firstly, we considered relatively few images compared to other multi-view approaches. While this keeps the capture process simple, it is probably not the best option for Internet images where the abundance of data can be exploited more explicitly—or may even be explicitly needed. Secondly, the need for an example object is the biggest hindrance. Conceptually, it would seem obvious to combine the presented approach with ideas from Chapter 6 and use a partial reconstruction to replace the reference object. We do, however, fear that the reference profiles would contain too many errors. This would then require to allow for more deviation during matching which again increases the chance of false matches. Given that multi-view photometric stereo is not yet a solved problem on laboratory data, the presented approach gives, nevertheless, an important impulse.

The most important insight gained is that surface orientation can be recovered relatively well even at incorrect depth. This enabled us to constrain the geometry estimation which otherwise is very unstable. We believe that this finding can be exploited in broader contexts as well and has potential to help overcome problems with textureless surface regions in other algorithms. Experimenting with objects that contain textured as well as untextured areas would be interesting and could lead to hybrid approaches.

Future Work: On the technical side, the most pressing task for future work is speeding up the matching process. This is the performance bottleneck in our current implementation and makes experiments tedious because the overall optimization then lasts several hours. From a conceptual point of view, finding ways to cope with depth discontinuities would be beneficial. These are not handled yet by our approach, but might be addressed through spatially varying weights β of the coupling term.

How to choose these weights, *e.g.* based on gradients in the master image, has to be investigated. Another strategy could be to employ different loss functions in the energy formulation. This change is straightforward in our framework, but needs careful and extensive evaluation.

Further ideas for future work arise when examining the connections to other approaches on an abstract level. Once the proxy geometry is obtained, Hernandez *et al.* [Hernandez08], Lim *et al.* [Lim05], or Zhang *et al.* [Zhang12] iterate between two optimization steps. The first one finds optimal normals based on the current depth estimate and associated image intensities. The second one updates the geometry based on these normal estimates:

$$n^i = \arg\min_n E_1(x^i, n), \qquad x^{i+1} = \arg\min_x E_2(x, n^i). \tag{7.10}$$

In our case, the optimization is not split into two steps which induce a certain search pattern, instead, the solver can vary depth and normals at the same time. It might be worth experimenting with other optimization strategies to achieve faster convergence and better avoidance of local minima. Also, a combination of appearance-based matching with the commonly used mesh representation would be interesting. While this complicates parallel processing, it would provide a more explicit occlusion handling and thus allow us to set more rigorous thresholds in the robust matching step. Finally, triangles provide a constraint on depth that usually spans several neighboring pixels. Our coupling term only considers pairs of two. Adding a constraint on a larger scale might prevent the failure case we observed for one view of the *diffuse owl*.

Chapter 8

Conclusion

We conclude this thesis with a summary of our contributions and a discussion of the insights gained. Finally, we give an outlook on future work.

8.1 Summary

We first present a radiometric camera model based on image meta information and evaluate its performance on Internet data. Compared to approaches for calibration in the laboratory, a ground truth measurement is impossible. We therefore use the Moon as a reference target and find that absolute luminance can be recovered up to plus/minus one f-stop. We also find that this level of accuracy allows plausible predictions of perceptual effects under low light conditions, *e.g.* decreased visual acuity and loss of color, from regular color images.

For photometric stereo under controlled conditions, one of the first steps is to calibrate the light source. We propose a novel technique to recover the position and not just the direction of a point light source close to the scene. The idea is to minimize the image space error of highlights generated by reflections of the light source on a spherical surface. In contrast to the traditional approach of minimizing ray distances, this error takes the perspective camera model into account. We also present an additional technique based on direct triangulation which does not rely on reflections. In an extensive evaluation, we find that the reflection-based approaches are accurate in the order of centimeters and that the direct triangulation achieves even millimeter accuracy.

Before focusing on orientation reconstruction in uncontrolled scenarios, it is important to know how well a carefully calibrated setup performs in order to put the results into perspective. We experimentally derive uncertainty estimates for the light source directions and image intensities. In an error analysis, we find that these do not allow a better reconstruction quality than 1° for our setup, which is only slightly below the experimental results of about 2.5°.

We then introduce three novel photometric reconstruction techniques that remove several restrictions from previous works. Most notably, we present the first approach to recover the full surface orientation and reflectance of objects captured by an outdoor webcam on the Internet. The key idea lies in the combination of an outdoor lighting model and a parametric material representation. Furthermore, the success is also attributed to the image selection scheme which extracts those images from the huge

amount of input that are well suited for reconstruction. The final result can be used to create renderings under novel lighting conditions.

Next, we study whether reconstructions on uncontrolled data are also possible without a parametric representation and explicit calibration. Example-based photometric stereo techniques promise a good generality but are rarely considered because of their need for a reference object. Our idea is to replace the reference with a partial reconstruction obtained purely from images of the target object. Such an imperfect reference object leads to new challenges but allows us to recover the surface orientation independent of a specific lighting model. This approach also demonstrates how information from two complementary techniques, multi-view stereo and photometric stereo, can be fused to overcome their individual weaknesses.

Finally, we present an approach that removes another restriction from photometric stereo: it no longer assumes a fixed camera. In contrast to other works, we do not rely on silhouettes or an initial multi-view stereo reconstruction to provide the correspondences but base all steps on shading information generated by varying illumination. We also employ a per-view energy formulation which can easily be extended with well-studied image-based regularization terms. Our formulation exploits the important insight that, for textureless objects, normals are reconstructed more reliably than depth and couples both components to obtain a consistent surface. Again, this approach avoids any light calibration and handles non-Lambertian reflectance.

8.2 Discussion

One of the motivations for this thesis was to explore the limits of photometric reconstruction methods in terms of input data and capture setup. A lot of research is concerned with finding novel aspects of image formation that can be exploited for photometric stereo or with increasing the performance of existing techniques. Many of these developments are still close to the first approaches 35 years ago regarding their restrictions on input data. Only few techniques focus on reducing the overhead in the capture setup and calibration or even think about Internet images as a data source.

Two main challenges for a wide-spread application are the calibration of light and, especially, camera response. In a pipeline from images to 3D model, these have to be considered along with the actual reconstruction algorithm. We have presented ways to either simplify the task through additional assumptions about outdoor illumination or to avoid calibration as much as possible using some form of reference geometry in the scene. Another challenge that has got some attention in recent years is extending photometric stereo for multi-view settings. Any such technique has to deal with missing pixel correspondences between images. This challenge can be addressed either by assuming that a suitable proxy geometry is available or by jointly estimating depth and normals. We have shown that the latter is possible for textureless objects and demonstrated how the reconstruction can be achieved without explicit light calibration.

Putting a focus on Internet images ensures that the input data is as uncontrolled as possible and forces us to think about requirements a lot more carefully than in traditional approaches. While this poses interesting research challenges, *e.g.* selecting suitable subsets of images, it also raises the question whether we actually need

reconstructions from Internet images. It is true that if a user wants to reconstruct his personal environment, the setting is usually much more controlled than arbitrary Internet data. For example, the radiometric response could be calibrated beforehand, and if the flash is used as light source, its position in camera coordinates is fixed and already known. We believe, however, that investigating the more general case is worthwhile because insights can be more easily transferred. An example from this thesis is the replacement of a reference object with a partial reconstruction which is an interesting option in both cases. Besides, more and more parts of the world will be covered more densely by images and videos in the future, allowing reconstructions on a scale that would not be possible for an individual.

In conclusion, this thesis presents a major step towards photometric reconstructions on uncontrolled data. We have developed algorithms to overcome challenges such as non-Lambertian reflectance, unknown camera response, and unknown illumination. One insight is that solving these problems on Internet data is more difficult than simply adapting known techniques from the laboratory. Especially the calibration is a crucial difference to traditional approaches. In addition, the low quality in terms of spatial resolution, compression artifacts, incorrect meta data, *etc.*, make reconstructions even more challenging. A second insight is that the amount of input data provided by Internet images is an advantage and a disadvantage at the same time. There is a good chance that images suitable for photometry are available in the input set but on the other hand it contains a lot of outliers. This demands an image selection geared towards this specific type of reconstruction, which was not considered previously. Finally, we have found that even on imperfect data a plausible simulation of novel impressions, *e.g.* relighting webcam images or creating low light perceptual effects, is possible. That hints at interesting questions relating image-based reconstructions and human perception which we leave for future work.

8.3 Future Work

Returning to the introductory problem statement, we repeat the ultimate goal of computer vision as described by Horn

> "A truly general-purpose vision system would have to deal with all aspects of vision and be applicable to all problems that can be solved using visual information." (B.K.P. Horn [Horn86a])

and ask how close we have come with respect to shape and reflectance reconstructions. We have definitely extended the space of input data and application scenarios that had been considered for photometric techniques. For Internet webcams, we can recover surface orientation and reflectance even for non-Lambertian surfaces. In the laboratory, we reconstruct objects of complex reflectance and unknown illumination from multiple views. Thus, to reach an even more general system, the next steps are to remove the single-view restriction from the first technique or to remove the example object from the second one.

For such an algorithm, community photo collections which contain images from different camera models, dramatic variation in scale, *etc.* could be the next challenge in terms of "all problems that can be solved using visual information". It is, however, unclear if a solution actually exists in that case. Another direction could be to keep

using particular subsets of Internet data, *e.g.* webcams as we did, and try to increase the robustness and accuracy. Research has just begun to consider Internet images for photometry, and future algorithms will probably extend it in both directions.

A photometric approach that reconstructs a detailed geometric model and reflectance properties on community photo collections is not yet known. Haber *et al.* [Haber09] present a first step but do not exploit the information encoded in the normals to recover detailed geometry. Thus, there is definitely a lot of room for improvement. Given our experiences with multi-view data on the one hand and Internet photometry on the other, we would suggest to approach this problem with a pixel or region selection scheme in addition to an image selection. If we could classify pixels according to their diffuseness, *cf.* Shen and Tan [Shen09], before running the actual reconstruction—and without knowing the camera response—we could first perform a light and response calibration similar to Diaz and Sturm [Diaz11a] only on the (almost) diffuse ones. In a second step, a full non-Lambertian reconstruction of all object regions could then recover normals and material properties. It is, however, unclear, if illumination recovered from the diffuse parts is sufficient in accuracy and frequency range to reconstruct specular materials.

In general, bringing the image creation models closer to reality, *e.g.* incorporating interreflections, while still being able to robustly invert them would be desirable on the path towards a general vision system. We do, however, not believe that such a truly general system or reconstruction technique can be built on a single algorithm or key idea. Even for different techniques from the same general class, *e.g.* multi-view stereo, it often depends on the scene which one outperforms the others. Thus, combinations of different approaches seem like a promising direction. We have mentioned the complementary properties of photometric stereo and multi-view stereo several times, but only the technique in Chapter 6 actually exploits information from both. Combining these more tightly than only using one to initialize the other could help to better leverage the advantages of both. More generally thinking, a combination of several methods with different advantages and disadvantages would be very interesting. If we had multiple techniques at our proposal, the interesting challenges would then be to find ways of combining their results or to automatically decide which one was better suited for a particular pixel or scene region. These questions also provide interconnections to topics such as higher-level scene understanding and model selection.

Focusing on Internet data raises the question of what quality we can expect from any reconstruction technique given such images. We have shown examples, but a general answer whether better results are at all possible is still an open question. Since ground truth data for Internet scenes is usually not available, a quantitative evaluation or the comparison of several techniques in a benchmark are difficult. Many works rely on qualitative results and only perform comparisons for controlled data. We exploit simple geometric primitives such as a cylinder or 90° edge to provide an intuition of the accuracy on webcam datasets. Better ways to evaluate results on Internet data would help the development of this field tremendously. Acquiring ground truth geometry, reflectance, and illumination for interesting scenes over a long period of time is certainly a huge effort but should be considered in the future.

Evaluation also raises the questions of a good error measure and how accurate we have to be—especially since the goal set by Horn does not define a desired level of accuracy. Like many others, we have mostly concentrated on the angular error

of normal maps. But a reconstruction is often part of a larger pipeline, and looking beyond the raw results might be appropriate. For example, the normals of the *tower* in Chapter 5 are slightly flattened in comparison with the ground truth, but this is hardly noticeable in the relighting results. Returning to one of our motivations, we ask whether the reconstructions obtained are suitable to (re-)create impressions of an actual observer. Currently, we can reproduce plausible *images* but there is still a lot of work to do until we can reproduce the *perception* a user has of a scene. Quantifying this perceptual component is not only interesting with respect to the results presented here but also for any other technique, *e.g.* multi-view stereo in combination with image-based rendering [Goesele10a].

Finally, research towards more general reconstruction algorithms has to consider an even larger spectrum of uncontrolled scenarios. We have used Internet images which are not only uncontrolled but often contain non-ideal data due to heavy post-processing, down-scaling, and the fact that they were not intended for reconstruction. Data sources that are better behaved but still uncontrolled, *e.g.* images taken by a single user with a reconstruction goal in mind, avoid these problems and could still provide valuable insights. Finding the right balance between expecting too much in terms of generality and still providing an inspiring challenge is therefore important for the progress of this field.

Bibliography

Publications (co-)authored by Jens Ackermann

[Ackermann09] J. Ackermann, P. Baecher, T. Franzel, M. Goesele, and K. Hamacher. Massively-parallel simulation of biochemical systems. In *Proceedings of Massively Parallel Computational Biology on GPUs (BioGPU2009)*, Lecture Notes in Informatics. 2009.

[Ackermann10] J. Ackermann, M. Ritz, A. Stork, and M. Goesele. Removing the example from example-based photometric stereo. In *European Conference on Computer Vision Workshops*, Lecture Notes in Computer Science. 2010.

[Ackermann12] J. Ackermann, F. Langguth, S. Fuhrmann, and M. Goesele. Photometric stereo for outdoor webcams. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Ackermann13a] J. Ackermann, S. Fuhrmann, and M. Goesele. Geometric point light source calibration. In *Proceedings of the Vision, Modeling, and Visualization Workshop*. 2013.

[Ackermann13b] J. Ackermann and M. Goesele. How bright is the Moon? Recovering and using absolute luminance values from internet images. In *Computational Color Imaging - 4th International Workshop*, Lecture Notes in Computer Science. 2013.

[Ackermann14] J. Ackermann, F. Langguth, S. Fuhrmann, A. Kuijper, and M. Goesele. Multi-view photometric stereo by example. In *Proceedings of the International Conference on 3D Vision*. 2014.

[Beljan12] M. Beljan, J. Ackermann, and M. Goesele. Consensus multi-view photometric stereo. In *DAGM/OAGM Symposium*, Lecture Notes in Computer Science. 2012.

[Fuhrmann10] S. Fuhrmann, J. Ackermann, T. Kalbe, and M. Goesele. Direct resampling for isotropic surface remeshing. In *Proceedings of the Vision, Modeling, and Visualization Workshop*. 2010.

[Goesele10a] M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klowsky, D. Steedly, and R. Szeliski. Ambient point clouds for view interpolation. *ACM Transactions on Graphics*, 2010.

[Goesele10b] M. Goesele, J. Ackermann, S. Fuhrmann, R. Klowsky, F. Langguth, P. Muecke, and M. Ritz. Scene reconstruction from community photo collections. *IEEE Computer*, 2010. (invited article).

List of Advised Theses

[Beljan11] M. Beljan. *Multi-View Photometric Stereo Using a Normal Consistency Approach*. Master's thesis, TU Darmstadt, 2011.

[Franek13] A. Franek. *Enhancing GPS Precision using Structure from Motion*. Master's thesis, TU Darmstadt, 2013. (co-supervised with Simon Fuhrmann).

[Langguth10] F. Langguth. *Photometric Stereo for Outdoor Webcams*. Bachelor's thesis, TU Darmstadt, 2010.

[Ritz09] M. Ritz. *A Combined Multi-View Stereo and Photometric Stereo Approach*. Master's thesis, TU Darmstadt, 2009.

References

[Abrams12] A. Abrams, C. Hawley, and R. Pless. Heliometric stereo: Shape from sun position. In *European Conference on Computer Vision*, pages 357–370. 2012.

[Adams83] A. Adams. *Examples: the making of 40 photographs*. Little, Brown Boston, 1983.

[Agarwal] S. Agarwal, K. Mierle, and Others. Ceres solver. `code.google.com/p/ceres-solver`. Accessed 2014-04-17.

[Agarwal09] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *IEEE International Conference on Computer Vision*, pages 72–79. 2009.

[Agrawal06] A. K. Agrawal, R. Raskar, and R. Chellappa. What is the range of surface reconstructions from a gradient field? In *European Conference on Computer Vision*, pages 578–591. 2006.

[Ahmed08] N. Ahmed, C. Theobalt, P. Dobrev, H.-P. Seidel, and S. Thrun. Robust fusion of dynamic shape and normal capture for high-quality reconstruction of time-varying geometry. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Aittala13] M. Aittala, T. Weyrich, and J. Lehtinen. Practical SVBRDF capture in the frequency domain. *ACM Transactions on Graphics*, 32:110:1–110:12, 2013.

[Alexander10] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, and P. Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30:20–31, 2010.

[Alldrin07a] N. G. Alldrin and D. J. Kriegman. Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach. In *IEEE International Conference on Computer Vision*, pages 1–8. 2007.

[Alldrin07b] N. G. Alldrin, S. P. Mallick, and D. J. Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

[Alldrin08] N. G. Alldrin, T. E. Zickler, and D. J. Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Aoto12] T. Aoto, T. Taketomi, T. Sato, Y. Mukaigawa, and N. Yokoya. Position estimation of near point light sources using a clear hollow sphere. In *International Conference on Pattern Recognition*, pages 3721–3724. 2012.

[Barron12] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Barron13] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Basri01a] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. In *IEEE International Conference on Computer Vision*. 2001.

[Basri01b] R. Basri and D. Jacobs. Photometric stereo with general, unknown lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2001.

[Bayer76] B. B. Bayer. Color image array, 1976. United States Patent 3971065.

[Belhumeur98] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28:245–260, 1998.

[Belhumeur99] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35:33–44, 1999.

[Bell13] S. Bell, P. Upchurch, N. Snavely, and K. Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics*, 32:111:1–111:17, 2013.

[Bendels05] G. H. Bendels, R. Schnabel, and R. Klein. Detail-preserving surface inpainting. In *The 6th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, pages 41–48. 2005.

[Bertero88] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889, 1988.

[Birkbeck06] N. Birkbeck, D. Cobzas, M. Jagersand, and P. Sturm. Variational shape and reflectance estimation under changing light and viewpoints. In *European Conference on Computer Vision*. 2006.

[Blinn77] J. F. Blinn. Models of light reflection for computer synthesized pictures. *SIGGRAPH Computer Graphics*, 11:192–198, 1977.

[Blostein89] D. Blostein and N. Ahuja. Shape from texture: Integrating texutre-element extraction and surface estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1233–1251, 1989.

[Bonfort03] T. Bonfort and P. Sturm. Voxel carving for specular surfaces. In *IEEE International Conference on Computer Vision*. 2003.

[Boykov03] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. In *IEEE International Conference on Computer Vision*, pages 26–33. 2003.

[Brady09] M. Brady and G. E. Legge. Camera calibration for natural image studies and vision research. *Journal of the Optical Society of America*, 26:30–42, 2009.

[Bronstein08] I. Bronstein and K. Semendjajew. *Taschenbuch der Mathematik*. Harri Deutsch Verlag, 2008.

[Brunger93] A. P. Brunger and F. C. Hooper. Anisotropic sky radiance model based on narrow field of view measurements of shortwave radiance. *Solar Energy*, 51:53–64, 1993.

[Buratti96] B. J. Buratti, J. K. Hillier, and M. Wang. The lunar opposition surge: Observations by Clementine. *Icarus*, 124:490–499, 1996.

[Burt80] P. Burt and B. Julesz. A disparity gradient limit for binocular fusion. *Science*, 208:615–617, 1980.

[Carceroni01] R. L. Carceroni and K. N. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape & reflectance. In *IEEE International Conference on Computer Vision*. 2001.

[Chakrabarti09] A. Chakrabarti, D. Scharstein, and T. E. Zickler. An empirical camera model for internet color vision. In *British Machine Vision Conference*, pages 51.1–51.11. 2009.

[Chandraker05] M. K. Chandraker, F. Kahl, and D. J. Kriegman. Reflection on the generalized bas-relief ambiguity. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2005.

[Chandraker11] M. K. Chandraker, J. Bai, and R. Ramamoorthi. A theory of differential photometric stereo for unknown BRDFs. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2011.

[Chandraker13] M. K. Chandraker, D. Reddy, Y. Wang, and R. Ramamoorthi. What object motion reveals about shape with unknown BRDF and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Chen06] T. Chen, M. Goesele, and H.-P. Seidel. Mesostructure from specularity. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2006.

[Coleman82] E. N. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing*, 18:309–328, 1982.

[Curless96] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of ACM SIGGRAPH*, pages 303–312. 1996.

[Darula02] S. Darula and R. Kittler. CIE general sky standard defining luminance distributions. In *Proceedings eSim*. 2002.

[Daum98] M. Daum and G. Dudek. On 3-D surface reconstruction using shape from shadows. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1998.

[deAguiar10] E. de Aguiar, L. Sigal, A. Treuille, and J. K. Hodgins. Stable spaces for real-time clothing. *ACM Transactions on Graphics*, 29:106:1–106:9, 2010.

[Debevec97] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of ACM SIGGRAPH*. 1997.

[Debevec00] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of ACM SIGGRAPH*. 2000.

[Delaunoy10] A. Delaunoy, E. Prados, and P. N. Belhumeur. Towards full 3D helmholtz stereovision algorithms. In *12th Asian Conference on Computer Vision*. 2010.

[Diaz11a] M. Diaz and P. Sturm. Exploiting image collections for recovering photometric properties. In *Computer Analysis of Images and Patterns*. 2011.

[Diaz11b] M. Diaz and P. Sturm. Radiometric calibration using photo collections. In *IEEE International Conference on Computational Photography*. 2011.

[Dosselmann13] R. Dosselmann and X. D. Yang. Improved method of finding the illuminant direction of a sphere. *SPIE Journal of Electronic Imaging*, 22, 2013.

[Drbohlav05] O. Drbohlav and M. Chantler. On optimal light configurations in photometric stereo. In *IEEE International Conference on Computer Vision*. 2005.

[Durand00] F. Durand and J. Dorsey. Interactive tone mapping. In *Eurographics Workshop on Rendering Techniques*. 2000.

[Durou09] J.-D. Durou, J.-F. Aujol, and F. Courteille. Integrating the normal field of a surface in the presence of discontinuities. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 261–273. 2009.

[Eberly] D. Eberly. Computing a point of reflection on a sphere. `www.geometrictools.com`. Accessed 2014-04-17.

[Ellis66] D. Ellis. Illumination received from the moon. *Journal of the Royal Astronomical Society of Canada*, 60:221–224, 1966.

[Eriksson10] A. Eriksson and A. van den Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Favaro12] P. Favaro and T. Papadhimitri. A closed-form solution to uncalibrated photometric stereo via diffuse maxima. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Ferwerda96] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg. A model of visual adaptation for realistic image synthesis. In *Proceedings of ACM SIGGRAPH*, pages 249–258. 1996.

[Fleming03] R. W. Fleming, R. O. Dror, and E. H. Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3:347–368, 2003.

[Frahm05] J.-M. Frahm, K. Koeser, D. Grest, and R. Koch. Markerless augmented reality with light source estimation for direct illumination. In *Conference on Visual Media Production*. 2005.

[Frahm10] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *European Conference on Computer Vision*. 2010.

[Frankot88] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:439–451, 1988.

[Fuhrmann11] S. Fuhrmann and M. Goesele. Fusion of depth maps with multiple scales. *ACM Transactions on Graphics*, 30:148:1–148:8, 2011.

[Furukawa10a] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Furukawa10b] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1362–1376, 2010.

[Garg09] R. Garg, H. Du, S. M. Seitz, and N. Snavely. The dimensionality of scene appearance. In *IEEE International Conference on Computer Vision*. 2009.

[Georghiades03] A. S. Georghiades. Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In *IEEE International Conference on Computer Vision*. 2003.

[Ghosh11] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics*, 30:129:1–129:10, 2011.

[Goesele07] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision*. 2007.

[Goldman05] D. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and spatially-varying BRDFs from photometric stereo. In *IEEE International Conference on Computer Vision*, pages 341–348. 2005.

[Goral84] C. M. Goral, K. E. Torrance, D. P. Greenberg, and B. Battaile. Modeling the interaction of light between diffuse surfaces. In *Proceedings of ACM SIGGRAPH*. 1984.

[Green92] D. W. E. Green. Correcting for atmospheric extinction. *International Comet Quarterly*, 14:55–59, 1992.

[Grossberg03] M. D. Grossberg and S. K. Nayar. What is the space of camera response functions? In *IEEE Conference on Computer Vision and Pattern Recognition*. 2003.

[Grossberg04] M. Grossberg and S. Nayar. Modeling the space of camera response functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1272–1282, 2004.

[Haber09] T. Haber, C. Fuchs, P. Bekaert, H.-P. Seidel, M. Goesele, and H. P. Lensch. Relighting objects from image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–634. 2009.

[Han13] Y. Han, J.-Y. Lee, and I. S. Kweon. High quality shape from a single RGBD image under uncalibrated natural illumination. In *IEEE International Conference on Computer Vision*. 2013.

[Hapke66] B. Hapke. An improved theoretical lunar photometric function. *The Astronomical Journal*, 71:333–339, 1966.

[Hara05] K. Hara, K. Nishino, and K. Ikeuchi. Light source position and reflectance estimation from a single view without the distant illumination assumption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:493–505, 2005.

[Hartley97] R. I. Hartley and P. F. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997.

[Hartley06] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2006.

[Hartt89] K. Hartt and M. Carlotto. A method for shape-from-shading using multiple images acquired under different viewing and lighting conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1989.

[Hayakawa94] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of the Optical Society of America*, 11(11):3079–3089, 1994.

[Healey86] G. Healey and T. O. Binford. Local shape from specularity. Technical report, Stanford University, 1986.

[Hernandez08] C. Hernandez, G. Vogiatzis, and R. Cipolla. Multi-view photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:548–554, 2008.

[Hernitschek08] N. Hernitschek, E. Schmidt, and M. Vollmer. Lunar eclipse photometry: absolute luminance measurements and modeling. *Applied optics*, 47(34):62–71, 2008. ISSN 1539-4522.

[Hertzmann03] A. Hertzmann and S. Seitz. Shape and materials by example: a photometric stereo approach. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2003.

[Hertzmann05] A. Hertzmann and S. M. Seitz. Example-based photometric stereo: shape reconstruction with general, varying BRDFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1254–1264, 2005. ISSN 0162-8828.

[Higo10] T. Higo, Y. Matsushita, and K. Ikeuchi. Consensus photometric stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Holroyd08] M. Holroyd, J. Lawrence, G. Humphreys, and T. E. Zickler. A photometric approach for estimating normals and tangents. *ACM Transactions on Graphics*, 27:133:1–133:9, 2008.

[Holroyd10] M. Holroyd, J. Lawrence, and T. E. Zickler. A coaxial optical scanner for synchronous acquisition of 3D geometry and surface reflectance. *ACM Transactions on Graphics*, 29:99:1–99:12, 2010.

[Horn70] B. K. P. Horn. Shape from shading: a method for obtaining the shape of a smooth opaque object from one view. Technical report, MIT Artificial Intelligence Laboratory, 1970.

[Horn78] B. K. P. Horn, R. J. Woodham, and W. M. Silver. Determining shape and reflectance using multiple images. Technical report, MIT Artificial Intelligence Laboratory, 1978.

[Horn86a] B. K. P. Horn. *Robot Vision*. MIT Press, 1986.

[Horn86b] B. K. P. Horn and M. J. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33:174–208, 1986.

[Horovitz04] I. Horovitz and N. Kiryati. Depth from gradient fields and control points: Bias correction in photometric stereo. *Image and Vision Computing*, 22:681–694, 2004.

[Ikehata12] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa. Robust photometric stereo using sparse regression. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[ISO06] ISO. 12232: Photography - digital still cameras - determination of exposure index, ISO speed ratings, standard output sensitivity, and recommended exposure index, 2006.

[Jacobs07a] N. Jacobs, N. Roman, and R. Pless. Consistent temporal variations in many outdoor scenes. 2007.

[Jacobs07b] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *IEEE International Conference on Computer Vision*. 2007.

[Jacobs09] N. Jacobs, W. Burgin, R. Speyer, D. Ross, and R. Pless. Adventures in archiving and using three years of webcam images. In *Computer Vision and Pattern Recognition Workshops*, pages 39–46. 2009.

[Jacobs10] N. Jacobs, B. Bies, and R. Pless. Using cloud shadows to infer scene structure and camera calibration. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Jacobs13a] N. Jacobs, A. Abrams, and R. Pless. Two cloud-based cues for estimating scene structure and camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:2526–2538, 2013.

[Jacobs13b] N. Jacobs, M. T. Islam, and S. Workman. Cloud motion as a calibration cue. In *CVPR*. 2013.

[Jensen01] H. W. Jensen, F. Durand, M. M. Stark, P. Shirley, J. Dorsey, and S. Premoze. A physically-based night sky model. In *Proceedings of ACM SIGGRAPH*. 2001.

[Jin03] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo beyond lambert. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2003.

[Jin05] H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 63:175–189, 2005.

[Johnson11] M. K. Johnson and E. H. Adelson. Shape estimation in natural illumination. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2011.

[Joshi07] N. Joshi and D. Kriegman. Shape from varying illumination and viewpoint. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

[Kanbara04] M. Kanbara and N. Yokoya. Real-time estimation of light source environment for photorealistic augmented reality. In *International Conference on Pattern Recognition*. 2004.

[Kazhdan06] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, pages 61–70. 2006.

[Kieffer05] H. H. Kieffer and T. C. Stone. The spectral irradiance of the moon. *The Astronomical Journal*, 129(6):2887–2901, 2005.

[Kim08a] S. J. Kim, J.-M. Frahm, and M. Pollefeys. Radiometric calibration with illumination change for outdoor scene analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Kim08b] S. J. Kim and M. Pollefeys. Robust radiometric calibration and vignetting correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:562–576, 2008.

[Kim12] S. J. Kim, H. T. Lin, Z. Lu, S. Suesstrunk, S. Lin, and M. S. Brown. A new in-camera imaging model for color computer vision and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:2289–2302, 2012.

[Klette96] R. Klette and K. Schluens. Height data from gradient fields. In *Machine Vision Applications, Architectures, and Systems Integration V*, pages 204–215. 1996.

[Koppal06] S. J. Koppal and S. G. Narasimhan. Clustering appearance for scene analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1323–1330. 2006.

[Krawczyk05] G. Krawczyk, K. Myszkowski, and H.-P. Seidel. Perceptual effects in real-time tone mapping. In *21st Spring Conference on Computer Graphics*. 2005.

[Kriegman01] D. J. Kriegman and P. N. Belhumeur. What shadows reveal about object structure. *Journal of the Optical Society of America*, 18:1804–1813, 2001.

[Kuthirummal08] S. Kuthirummal, A. Agarwala, D. B. Goldman, and S. K. Nayar. Priors for large photo collections and what they reveal about cameras. In *European Conference on Computer Vision*, pages 74–87. 2008.

[Laffont12a] P.-Y. Laffont, A. Bousseau, and G. Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Transactions on Visualization and Computer Graphics*, 19:210–224, 2012.

[Laffont12b] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics*, 31:202:1–202:11, 2012.

[Lalonde08] J. Lalonde, S. Narasimhan, and A. Efros. What does the sky tell us about the camera? In *European Conference on Computer Vision*, pages 354–367. 2008.

[Lalonde09] J. Lalonde, A. Efros, and S. G. Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *ACM Transactions on Graphics*, 28(5):131:1–131:10, 2009.

[Lalonde10] J. Lalonde, S. G. Narasimhan, and A. Efros. What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 88:24–51, 2010.

[Lawrence06] J. Lawrence, A. Ben-Artzi, C. DeCoro, W. Matusik, H. Pfister, R. Ramamoorthi, and S. Rusinkiewicz. Inverse shade trees for non-parametric material representation and editing. *ACM Transactions on Graphics*, 25:735–745, 2006.

[Lee12] K. J. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. U. Lee, P. Tan, and S. Lin. Estimation of intrinsic image sequences from image+depth video. In *European Conference on Computer Vision*. 2012.

[Lempitsky07] V. S. Lempitsky and Y. Boykov. Global optimization for shape fitting. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

[Lensch00] H. P. A. Lensch, W. Heidrich, and H.-P. Seidel. Automated texture registration and stitching for real world models. In *Pacific Conference on Computer Graphics and Applications*. 2000.

[Lensch01] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatially varying materials. In *Eurographics Workshop on Rendering Techniques*. 2001.

[Lensch03] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Transactions on Graphics*, 22:234–257, 2003.

[Lim05] J. Lim, J. Ho, M.-H. Yang, and D. Kriegman. Passive photometric stereo from motion. In *IEEE International Conference on Computer Vision*. 2005.

[Lin04] S. Lin, J. Gu, S. Yamazaki, and H.-Y. Shum. Radiometric calibration from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2004.

[Litvinov05] A. Litvinov and Y. Y. Schechner. Addressing radiometric nonidealities: A unified framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 52–59. 2005.

[Lombardi12] S. Lombardi and K. Nishino. Reflectance and natural illumination from a single image. In *European Conference on Computer Vision*. 2012.

[Lu95] J. Lu and J. Little. Reflectance function estimation and shape recovery from image sequences of a rotating object. In *IEEE International Conference on Computer Vision*. 1995.

[Lu13] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Lucas81] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 121–130. 1981.

[Lun] Lunar phase angle. `the-moon.wikispaces.com/Phase`. Accessed 2014-04-17.

[Ma07] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Eurographics Symposium on Rendering Techniques*. 2007.

[Mann94] S. Mann and R. Picard. Being undigital with digital cameras: Extending dynamic range by combining differently exposed pictures. Technical Report 323, M.I.T. Media Lab Perceptual Computing Section, Boston, Massachusetts, 1994. Also appears, IS&T's 48th annual conference, Cambridge, Massachusetts, May 1995.

[Marschner98] S. R. Marschner. *Inverse Rendering for Computer Graphics*. Ph.D. thesis, Cornell University, 1998.

[Marschner00] S. R. Marschner, S. H. Westin, E. P. F. Lafortune, and K. E. Torrance. Image-based bidirectional reflectance distribution function measurement. *Applied Optics*, 39:2592–2600, 2000.

[Martinez-Verdu03] F. Martinez-Verdu, J. Pujol, M. Vilaseca, and P. Capilla. Characterization of a digital camera as an absolute tristimulus colorimeter. *Proceedings of the SPIE*, 47(4):279–295, 2003.

[Masselus02] V. Masselus, P. Dutré, and F. Anrys. The free form light stage. In *Eurographics Workshop on Rendering Techniques*. 2002.

[Matsushita07] Y. Matsushita and S. Lin. Radiometric calibration from noise distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

[Matusik03] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Transactions on Graphics*, 22:759–769, 2003.

[McNicholas34] H. J. McNicholas. Equipment for measuring the reflective and transmissive properties of diffusing media. *Journal of Research of the National Bureau of Standards*, 13:211–236, 1934.

[Mongkulmann11] W. Mongkulmann, T. Okabe, and Y. Sato. Photometric stereo with auto-radiometric calibration. In *International Conference on Computer Vision Workshops*. 2011.

[Mueller05] G. Mueller, G. H. Bendels, and R. Klein. Rapid synchronous acquisition of geometry and appearance of cultural heritage artefacts. In *The 6th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, pages 13–20. 2005.

[Murray-Coleman90] J. F. Murray-Coleman and A. M. Smith. The automated measurement of BRDFs and their application to luminaire modeling. *Journal of the Illuminating Engineering Society*, 19:87–99, 1990.

[Narasimhan02] S. G. Narasimhan, C. Wang, and S. K. Nayar. All the images of an outdoor scene. In *European Conference on Computer Vision*, pages 148–162. 2002.

[NASA] NASA. Horizons. `ssd.jpl.nasa.gov/?horizons`. Accessed 2014-04-17.

[Nayar89a] S. K. Nayar. Sphereo: Determining depth using two specular spheres and a single camera. In *Robotics Conferences*. 1989.

[Nayar89b] S. K. Nayar, K. Ikeuchi, and T. Kanade. Determining shape and reflectance of lambertian, specular, and hybrid surfaces using extended sources. In *International Workshop on Industrial Applications of Machine Intelligence and Vision*. 1989.

[Nayar90] S. K. Nayar, K. Ikeuchi, and T. Kanade. Shape from interreflections. Technical report, Carnegie Mellon University, 1990.

[Nayar93] S. K. Nayar, X.-S. Fang, and T. Boult. Removal of specularities using color and polarization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1993.

[Nehab05] D. Nehab, S. Rusinkiewicz, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics*, 24:536–543, 2005.

[Nehab08] D. Nehab, T. Weyrich, and S. Rusinkiewicz. Dense 3D reconstruction from specularity consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Nicodemus77] F. E. Nicodemus, J. C. Richmond, and J. J. Hsia. *Geometrical Considerations and Nomenclature for Reflectance*. U.S. Department of Commerce, 1977.

[NIST08] NIST. The international system of units (SI), 2008. U.S. Department of Commerce, National Institute of Standards and Technology.

[Nistér04] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:756–777, 2004.

[Okabe09] T. Okabe, I. Sato, and Y. Sato. Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In *IEEE International Conference on Computer Vision*. 2009.

[Okatani07] T. Okatani and K. Deguchi. On the wiberg algorithm for matrix factorization in the presence of missing components. *International Journal of Computer Vision*, 2007.

[Okatani11] T. Okatani, T. Yoshida, and K. Deguchi. Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms. In *IEEE International Conference on Computer Vision*. 2011.

[Okatani12] T. Okatani and K. Deguchi. Optimal integration of photometric and geometric surface measurements using inaccurate reflectance/illumination knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Oxholm12] G. Oxholm and K. Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision*. 2012.

[Panagopoulos11] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2011.

[Papadhimitri13] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Park13] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon. Multiview photometric stereo using planar mesh parameterization. In *IEEE International Conference on Computer Vision*. 2013.

[Paterson05] J. A. Paterson, D. Claus, and A. W. Fitzgibbon. BRDF and geometry capture from extended inhomogeneous samples using flash photography. In *Eurographics*. 2005.

[Pentland87] A. P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:523–531, 1987.

[Perez93] R. Perez, R. Seals, and J. Michalsky. All-weather model for sky luminance distribution-preliminary configuration and validation. *Solar Energy*, 50:235–245, 1993.

[Pharr12] M. Pharr and G. Humphries. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann Publishers, 2012.

[Phong75] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18:311–317, 1975.

[Powell01] M. W. Powell, S. Sarkar, and D. B. Goldgof. A simple strategy for calibrating the geometry of light sources. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[Preetham99] A. J. Preetham, P. Shirley, and B. Smits. A practical analytic model for daylight. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999.

[Press92] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*. 2nd edition, 1992.

[Ramamoorthi01] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse render. In *Proceedings of ACM SIGGRAPH*. 2001.

[Ray83] R. Ray, J. Birk, and R. B. Kelley. Error analysis of surface normals determined by radiometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:631–645, 1983.

[Reda04] I. Reda and A. Andreas. Solar position algorithm for solar radiation applications. *Solar Energy*, pages 577–589, 2004.

[Reinhard02] E. Reinhard, M. Stark, P. Shirley, and J. A. Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 2002.

[Rindfleisch65] T. Rindfleisch. A photometric method for deriving lunar topographic information. Technical report, California Insitute of Technology, 1965.

[Robertson99] M. A. Robertson, S. Borman, and R. L. Stevenson. Dynamic range improvement through multiple exposures. In *IEEE International Conference on Image Processing*, pages 159–163. 1999.

[Robertson03] M. A. Robertson, S. Borman, and R. L. Stevenson. Estimation-theoretic approach to dynamic range enhancement using multiple exposures. *Journal of Electronic Imaging*, 12:219–228, 2003.

[Romeiro10] F. Romeiro and T. E. Zickler. Blind reflectometry. In *European Conference on Computer Vision*. 2010.

[Ruiters09] R. Ruiters and R. Klein. Heightfield and spatially varying BRDF reconstruction for materials with interreflections. *Computer Graphics Forum*, 28:513–522, 2009.

[Ruiters12] R. Ruiters, C. Schwartz, and R. Klein. Data driven surface reflectance from sparse and irregular samples. *Computer Graphics Forum*, 31:315–324, 2012.

[Sato94a] Y. Sato and K. Ikeuchi. Reflectance analysis under solar illumination. Technical report, Carnegie Mellon University, 1994.

[Sato94b] Y. Sato and K. Ikeuchi. Temporal-color space analysis of reflection. *Journal of the Optical Society of America*, 11, 1994.

[Sato95] Y. Sato and K. Ikeuchi. Reflectance analysis under solar illumination. In *Proceedings of the Workshop on Physics-Based Modeling in Computer Vision*, pages 180–187. 1995.

[Sato99] I. Sato, Y. Sato, and K. Ikeuchi. Acquiring a radiance distribution to superimpose virtual objects onto a real scene. *Transactions on Visualization and Computer Graphics*, 1999.

[Sato07] I. Sato, T. Okabe, Q. Yu, and Y. Sato. Shape reconstruction based on similarity in radiance changes under varying illumination. In *IEEE International Conference on Computer Vision*, pages 1–8. 2007.

[Schuster10] V. Schuster. *BRDF Based Photo-Consistency*. Diploma thesis, Computer Graphics department, University of Bonn, 2010.

[Schwartz11] C. Schwartz, M. Weinmann, R. Ruiters, and R. Klein. Integrated high-quality acquisition of geometry and appearance for cultural heritage. In *The 12th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage*, pages 25–32. 2011.

[Seitz06] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 519–526. 2006. `vision.middlebury.edu/mview/`.

[Shafer84] S. A. Shafer. Using color to separate reflection components. Technical report, University of Rochester, 1984.

[Shen09] L. Shen and P. Tan. Photometric stereo and weather estimation using internet images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

[Shi10] B. Shi, Y. Matsushita, Y. Wei, C. Xu, and P. Tan. Self-calibrating photometric stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[Shi12a] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. A biquadratic reflectance model for radiometric image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Shi12b] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi. Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances. In *European Conference on Computer Vision*. 2012.

[Shlaer37] S. Shlaer. The relation between visual acuity and illumination. *The Journal of General Physiology*, 21(2):165–188, 1937.

[Silver80] W. M. Silver. *Determining Shape and Reflectance Using Multiple Images*. Master's thesis, Massachusetts Institute of Technology, 1980.

[Simakov03] D. Simakov, D. Frolova, and R. Basri. Dense shape reconstruction of a moving object under arbitrary, unknown lighting. In *IEEE International Conference on Computer Vision*. 2003.

[Simon08] I. Simon and S. M. Seitz. Scene segmentation using the wisdom of crowds. In *European Conference on Computer Vision*. 2008.

[Snavely06] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[Soatto03] S. Soatto, A. J. Yezzi, and H. Jin. Tales of shape and radiance in multi-view stereo. In *IEEE International Conference on Computer Vision*. 2003.

[Spencer95] G. Spencer, P. Shirley, K. Zimmerman, and D. P. Greenberg. Physically-based glare effects for digital images. In *Proceedings of ACM SIGGRAPH*, pages 325–334. 1995.

[Stokes] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta. A standard default color space for the internet - sRGB. `www.w3.org/Graphics/Color/sRGB.html`. Accessed 2014-04-17.

[Sunkavalli07] K. Sunkavalli, W. Matusik, H. Pfister, and S. Rusinkiewicz. Factored time-lapse video. *ACM Transactions on Graphics*, 2007.

[Sunkavalli08] K. Sunkavalli, F. Romeiro, W. Matusik, T. Zickler, and H. Pfister. What do color changes reveal about an outdoor scene? In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Sunkavalli10] K. Sunkavalli, T. E. Zickler, and H. Pfister. Visibility subspaces: Uncalibrated photometric stereo with shadows. In *European Conference on Computer Vision*. 2010.

[Tagare91] H. D. Tagare and R. J. P. deFigueiredo. A theory of photometric stereo for a class of diffuse non-lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 1991.

[Takai09] T. Takai, A. Maki, K. Niinuma, and T. Matsuyama. Difference sphere: An approach to near light source estimation. *Computer Vision and Image Understanding*, 2009.

[Tan05] R. T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:178–193, 2005.

[Tan07] P. Tan, S. P. Mallick, L. Quan, D. J. Kriegman, and T. E. Zickler. Isotropy, reciprocity and the generalized bas-relief ambiguity. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2007.

[Tan09] P. Tan and T. E. Zickler. A projective framework for radiometric image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2009.

[Tan11] P. Tan, L. Quan, and T. E. Zickler. The geometry of reflectance symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 2011.

[Tenenbaum00] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[Tompkin12] J. Tompkin, K. I. Kim, J. Kautz, and C. Theobalt. Videoscapes: Exploring sparse, unstructured video collections. *ACM Transactions on Graphics*, 2012.

[Torrance67] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America*, 57:1105–1112, 1967.

[Treuille04] A. Treuille, A. Hertzmann, and S. M. Seitz. Example-based stereo with general BRDFs. In *European Conference on Computer Vision*. 2004.

[Tunwattanapong13] B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. Debevec. Acquiring reflectance and shape from continuous spherical harmonic illumination. *ACM Transactions on Graphics*, 2013.

[UCL] UCL. Colour & vision research laboratory database. `www.cvrl.org/lumindex.htm`. Accessed 2014-04-17.

[Umeyama91] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

[USGS] USGS. Lunar calibration, robotic lunar observatory (ROLO). `www.moon-cal.org`. Accessed 2014-04-17.

[Verbiest08] F. Verbiest and L. V. Gool. Photometric stereo with coherent outlier handling and confidence estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

[Vlasic09] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popovic, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics*, 28, 2009.

[Vogiatzis06] G. Vogiatzis, C. Hernandez, and R. Cipolla. Reconstruction in the round using photometric normals and silhouettes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2006.

[Walthelm] A. Walthelm. Picturerelate. `www.walthelm.net/picture-relate/index.php`. Accessed 2014-04-17.

[Wang02] Y. Wang and D. Samaras. Estimation of multiple illuminants from a single image of arbitrary known geometry. In *European Conference on Computer Vision*. 2002.

[Ward92] G. J. Ward. Measuring and modeling anisotropic reflection. In *Proceedings of ACM SIGGRAPH*, pages 265–272. 1992.

[Ward97] G. J. Ward, H. Rushmeier, and C. Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. In *Proceedings of ACM SIGGRAPH*, volume 3. 1997.

[Weba] Webcam: Castle in bautzen. `www.budysin.de`.

[Webb] Webcam: Church in dresden. `www.frauenkirche.de`.

[Webc] Webcam: Roller coaster in rust. `www.europapark.de`.

[Webd] Webcam: Tower in mannheim. `www.mvv-energie.de`.

[Weber01] M. Weber and R. Cipolla. A practical method for estimation of point light-sources. In *British Machine Vision Conference*. 2001.

[Weber02] M. Weber, A. Blake, and R. Cipolla. Towards a complete dense geometric and photometric reconstruction under varying pose and illumination. In *British Machine Vision Conference*. 2002.

[Weinmann12] M. Weinmann, R. Ruiters, A. Osep, C. Schwartz, and R. Klein. Fusing structured light consistency and helmholtz normals for 3D reconstruction. In *British Machine Vision Conference*, pages 1–12. 2012.

[Weinmann13] M. Weinmann, A. Osep, R. Ruiters, and R. Klein. Multi-view normal field integration for 3D reconstruction of mirroring objects. In *IEEE International Conference on Computer Vision*, pages 2504–2511. 2013.

[Winnemoeller05] H. Winnemoeller, A. Mohan, J. Tumblin, and B. Gooch. Light waving: Estimating light positions from photographs alone. *Computer Graphics Forum*, 2005.

[Wong08] K.-Y. K. Wong, D. Schnieders, and S. Li. Recovering light directions and camera poses from a single sphere. In *European Conference on Computer Vision*. 2008.

[Woodham77] R. J. Woodham. *Reflectance Map Techniques for Analyzing Surface Defects in Metal Castings*. Ph.D. thesis, Massachusetts Institute of Technology, 1977.

[Woodham80] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, pages 139–144, 1980.

[Wu88] Z. Wu and L. Li. A line-integration based method for depth recovery from surface normals. *Computer Vision, Graphics, and Image Processing*, 43(1):53–66, 1988. ISSN 0734-189X.

[Wu06] T.-P. Wu and C.-K. Tang. Dense photometric stereo by expectation maximization. In *European Conference on Computer Vision*. 2006.

[Wu10] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *12th Asian Conference on Computer Vision*. 2010.

[Wu11] C. Wu, Y. Liu, Q. Dai, and B. Wilburn. Fusing multiview and photometric stereo for 3D reconstruction under uncalibrated illumination. *IEEE Transactions on Visualization and Computer Graphics*, 17, 2011.

[Wu12] C. Wu, K. Varanasi, and C. Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *European Conference on Computer Vision*. 2012.

[Wu13] Z. Wu and P. Tan. Calibrating photometric stereo by holistic reflectance symmetry analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Wueller07] D. Wueller and H. Gabele. The usage of digital cameras as luminance meters. *Proceedings of the SPIE*, 6502:1–11, 2007.

[Xiong12] Y. Xiong, K. Saenko, T. Darrell, and T. E. Zickler. From pixels to physics: Probabilistic color de-rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Xu08] S. Xu and A. Wallace. Recovering surface reflectance and multiple light locations and intensities from image data. *Pattern Recognition Letters*, 2008.

[Yang03] R. Yang, M. Pollefeys, and G. Welch. Dealing with textureless regions and specular highlights - a progressive space carving scheme using a novel photo-consistency measure. In *IEEE International Conference on Computer Vision*. 2003.

[Yoon10] K.-J. Yoon, E. Prados, and P. Sturm. Joint estimation of shape and reflectance using multiple images with known illumination conditions. *International Journal of Computer Vision*, 86:192–210, 2010.

[Yoshiyasu11] Y. Yoshiyasu and N. Yamazaki. Topology-adaptive multi-view photometric stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2011.

[Yu99] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of ACM SIGGRAPH*. 1999.

[Yu04] T. Yu, N. Xu, and N. Ahuja. Shape and view independent reflectance map from multiple views. In *European Conference on Computer Vision*. 2004.

[Yu13] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, D. Terzopoulos, and T. F. Chan. Outdoor photometric stereo. In *IEEE International Conference on Computational Photography*. 2013.

[Yuille97] A. Yuille and D. Snow. Shape and albedo from multiple images using integrability. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1997.

[Zhang99a] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, 1999.

[Zhang99b] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *IEEE International Conference on Computer Vision*. 1999.

[Zhang03] L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and motion under varying illumination: unifying structure from motion, photometric stereo, and multiview stereo. In *IEEE International Conference on Computer Vision*, pages 618–625. 2003.

[Zhang12] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu. Edge-preserving photometric stereo via depth fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2012.

[Zhou02] W. Zhou and C. Kambhamettu. Estimation of illuminant direction and intensity of multiple light sources. In *European Conference on Computer Vision*. 2002.

[Zhou04] W. Zhou and C. Kambhamettu. A unified framework for scene illuminant estimation. In *British Machine Vision Conference*. 2004.

[Zhou07] S. K. Zhou, G. Aggarwal, R. Chellappa, and D. W. Jacobs. Appearance characterization of linear lambertian objects, generalized photometric stereo and illumination-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 2007.

[Zhou13] Z. Zhou, Z. Wu, and P. Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[Zickler02] T. E. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. *International Journal of Computer Vision*, 49, 2002.

[Zickler06] T. E. Zickler. Reciprocal image features for uncalibrated helmholtz stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2006.

[Zinth98] W. Zinth and H. Körner. *Physik III: Optik, Quantenphänomene und Aufbau der Materie*. Oldenbourg, 1998.

Curriculum Vitae

Der Lebenslauf ist in der Online-Verison aus Gründen des Datenschutzes nicht enthalten.